

ENCYCLOPÉDIE
DES
SCIENCES
MATHEMATIQUES
GRANGE BATELIÈRE

LEXIQUE DE MATHÉMATIQUES

SUPPLÉMENT AU VOLUME XIII DE LA GRANDE ENCYCLOPÉDIE ALPHA DES SCIENCES ET DES TECHNIQUES

ABRÉVIATIONS

adj.	adjectif	*n.*	nom	*p. ex.*	par exemple
ex.	exemple	*par ext.*	par extension	*syn.*	synonyme

A

absorbant. *adj.* Un élément *a* d'un ensemble E muni d'une loi T est absorbant si, pour tout *b* appartenant à E, $a \top b = b \top a = a$.

accumulateur. *n.* Registre de l'U.A.L. (unité arithmétique et logique) dans lequel sont placés les résultats de la plupart des opérations élémentaires.

adhérent. *adj.* Tout point qui appartient à l'*adhérence* (ou fermeture) d'une partie A d'un espace topologique est dit point adhérent à A.

adresse. *n.* Numéro du registre mémoire ou mot mémoire contenant une information et permettant d'accéder à celle-ci.

aléatoire. *adj.* Soumis au hasard.

aléatoire. *adj.* **(fonction).** Se dit d'une fonction F telle que sa valeur F (*x*) dépend du résultat du tirage d'une ou plusieurs variables aléatoires.

aléatoire. *adj.* **(variable).** Soit $\mathcal{E}$ un espace des événements possibles. Soit X une application de $\mathcal{E}$ sur R, telle que l'image inverse par X de tout intervalle de R soit un élément de E, tribu associée à $\mathcal{E}$. X est dit variable aléatoire définie sur l'espace probabilisable ($\mathcal{E}$, E). On dit que X est une variable aléatoire purement *discrète* si la réalisation des événements non nuls de $\mathcal{E}$ ne peut conduire qu'à une série discrète de valeurs $x_n \in$ R, et absolument *continue* sur l'intervalle non nul $[a, b] \in$ R si aucun point de cet intervalle ne correspond à la réalisation d'un événement impossible.

algébrique. *adj.* **(entier).** Racine d'un polynôme à coefficients entiers, dont le coefficient du terme de plus fort degré est égal à 1. *Ex. :* $\sqrt{2}$ (racine de $x^2 - 2$).

algébrique. *adj.* **(extension).** Se dit d'une extension simple K (*l*) d'un corps K, telle qu'il existe au moins un polynôme non nul *f* (*x*) à coefficients dans K admettant *l* pour racine : *f* (*l*) = 0.

algébrique. *adj.* **(nombre).** Racine d'un polynôme à coefficients entiers. *Ex. :* $\sqrt{\frac{2}{3}}$ (racine de $3x^2 - 2$).

analytique. *adj.* **(fonction).** Se dit d'une fonction dont le développement en série de Taylor converge vers la valeur de la fonction (voir *holomorphe*).

anneau. *n.* Un ensemble A muni de deux lois de composition interne (une addition et une multiplication) est un *anneau* si :
1° A est un groupe abélien pour l'addition;
2° la multiplication est associative;
3° la multiplication est distributive par rapport à l'addition.

anneau intègre. Un anneau A est dit intègre ou anneau d'intégrité s'il n'existe pas d'éléments *x, y* de A différents tous deux de 0, tels que *xy* = 0.

application (d'un ensemble dans un autre ensemble). Relation telle que tout élément de l'ensemble de départ ait une image et une seule.

archimédien. *adj.* **(corps).** Qualifie un corps ordonné K dans lequel la propriété suivante est vérifiée : Pour tout couple (*a, b*) d'éléments de K strictement positifs, il existe un entier *n* tel que $n\,b \geqslant a$.

assembleur. *n.* Programme destiné à traduire en langage machine binaire un programme écrit en langage machine symbolique.

axiomatisation. *n.* La méthode axiomatique est au principe de toute mathématique rigoureuse. Bien avant d'avoir été employée avec une maîtrise parfaite dans les *Éléments* d'Euclide, elle a été abondamment décrite par Platon et par Aristote. Mais alors il ne s'agissait que de poser des énoncés primitifs et d'en dériver d'autres qui « consonnent » ou « s'harmonisent » avec eux. Autrement dit, il s'agissait d'une bonne approche de l'idée de déduction purement logique. Mais le caractère des énoncés primitifs était laissé dans l'ombre et prêtait à des confusions; on a beaucoup discuté sur le fait que ces énoncés étaient des vérités primitives au sens d'évidences premières accessibles à l'expérience immédiate, ou des conventions, ou des définitions. Depuis la fin du XIX[e] siècle, la généralisation de la méthode axiomatique a permis la mise en lumière de principes ensemblistes sous-jacents aux différentes disciplines mathématiques. Il a fallu axiomatiser la théorie des ensembles qu'on a ainsi trouvée au principe de toutes les mathématiques, ce qui conduit à la découverte des fameux paradoxes. L'application de la méthode axiomatique à la théorie des ensembles dut alors passer pour le principe des définitions implicites (on renonce à définir directement la notion d'ensemble implicitement définie par l'ensemble des axiomes) inhérent à la méthode axiomatique moderne inaugurée par Hilbert dans les *Fondements de la Géométrie* et caractérisée par le fait que les notions primitives, telles que point, droite, plan, sont des objets sans contenus intuitifs, définis pour les seules relations réciproques posées sous forme d'axiomes.

axiome de choix. L'« axiome de choix », qui doit son nom au mathématicien Zermelo, dit que si *t* est une famille d'ensembles disjoints et non vides, il est toujours possible de choisir dans chacun des membres de la famille un élément unique et de réunir tous les éléments ainsi distingués en un ensemble bien défini. En termes plus mathématiques, l'axiome énonce que le « produit cartésien » des membres de la famille *t* n'est pas vide, c'est-à-dire que parmi les sous-ensembles de l'ensemble $\bigcup t$ obtenu en réunissant tous les éléments de tous les membres de *t*, il y a au moins un sous-ensemble dont l'intersection avec chacun des membres de *t* est constituée par un élément unique.

B

base. *n.* Partie libre et génératrice d'un espace vectoriel. Tout espace vectoriel admet des bases.

base d'un système de numération. Nombre de symboles utilisés pour représenter un nombre dans ce système.

bidual. *n.* Soit E un espace normé et E' son dual topologique; on appelle bidual de E l'espace E'', dual topologique de E' muni de la topologie forte.

bijection. *n.* Application telle que son application réciproque est aussi une application.

bilinéaire. *adj.* **(forme).** Se dit d'une application de E × E (où E est un espace vectoriel) dans son corps de base K, linéaire par rapport au premier et au deuxième arguments.

binaire. *adj.* **(additionneur).** Se dit d'un circuit électronique réalisant l'addition en base 2.

binaire. *adj.* **(système).** Se dit d'un système de numération en base 2, c'est-à-dire utilisant les seuls symboles 0 et 1.

bistable. *n.* Voir *flip flop*.

bit. *n.* Chiffre binaire (de l'anglais *b*inary dig*it*); *par ext.* signal logique.

bon ordre. Un ensemble est muni d'un bon ordre si toute partie non vide de cet ensemble admet un plus petit élément.

bornée. *adj.* **(partie).** Qualifie une partie d'un espace métrique contenue dans une boule de rayon fini.

borne inférieure. On appelle borne inférieure d'une partie d'un ensemble ordonné le plus grand des *minorants* de cette partie.

borne supérieure. On appelle borne supérieure d'une partie d'un ensemble ordonné le plus petit des *majorants* de cette partie.

boucle. *n.* En informatique, séquence d'instructions exécutée plusieurs fois de suite.

boule. *n.* L'ensemble des points d'un espace métrique E, dont la distance à un point $a \in E$ est inférieure à un nombre $r > 0$ (respectivement inférieure ou égale) est dit boule ouverte (respectivement fermée) de centre a et de rayon r.

C

calcul infinitésimal. Inventé simultanément et indépendamment par Newton et Leibniz, le calcul infinitésimal a soulevé bien des difficultés avant que ne fût éclairci (au XIXe siècle) le concept de *limite,* puis rigoureusement défini celui de *nombre réel* (voir le chapitre *Histoire des Mathématiques*). Cette expression, aujourd'hui vieillie et remplacée par le terme *analyse* (héritage de l'expression *analyse infinitésimale* due à Euler et très répandue au siècle dernier), désigne l'ensemble des méthodes du calcul différentiel, du calcul intégral et du calcul des variations.

canonique. *adj.* **(analyse).** L'analyse canonique est une méthode statistique de traitement de données multidimensionnelles, qui consiste à étudier les relations entre deux groupes de variables.

caractéristique universelle. Ce concept leibnizien désigne l'art d'exposer les pensées au moyen d'un ensemble de signes et d'expressions qui en rendent possible et aisée la manipulation par les ressources du calcul.

cardinal. *n.* Deux ensembles ont même cardinal s'il existe une bijection de l'un sur l'autre. Pour un ensemble fini, le cardinal est le nombre d'éléments de cet ensemble.

catégorie. *n.* Un ensemble C muni d'une loi de composition interne (notée $\circ$) associative est une catégorie si :
1° il existe deux applications α et β de C dans une partie C_0 de C telles que la restriction de α et β à C_0 soit l'identité;
2° le composé $g \circ f$ de deux éléments de C est défini si et seulement si $\alpha(g) = \beta(f)$; on a alors :
$$\alpha(g \circ f) = \alpha(f) \text{ et } \beta(g \circ f) = \beta(g);$$
3° pour tout élément f de C, on a :
$$f \circ \alpha(f) = f = \beta(f) \circ f.$$

Cauchy (suite de). Si une suite (x_n) appartient à un espace métrique (E, d), elle est dite *de Cauchy* si pour tout $\varepsilon > 0$ il existe un entier p tel que dès que n et m sont des entiers supérieurs à p, on ait : $d(x_n, x_m) < \varepsilon$.

centrée. *adj.* **(variable aléatoire).** On dit que la variable aléatoire X est centrée si l'espérance mathématique de X est nulle : $E(X) = 0$.

cercle. *n.* Le cercle est, dans le plan, le lieu des points situés à une distance constante d'un point fixe. L'équation dans un système orthonormé d'un cercle de centre (α, β) et de rayon r est :
$$(x - \alpha)^2 + (y - \beta)^2 = r^2.$$

chaîne. *n.* Une combinaison linéaire à coefficients entiers de p-simplexes singuliers est appelée p-chaîne ou chaîne p-dimensionnelle.

chargeur. *n.* Programme destiné à charger en mémoire d'autres programmes. Parfois appelé *éditeur de liens.*

circulaire. *adj.* **(fonction).** Une fonction circulaire directe mesure un segment en fonction d'un angle, tandis qu'une fonction circulaire inverse mesure un angle en fonction d'un segment.

clan. *n.* Famille de parties d'un ensemble stable pour les opérations de réunion et de passage au complémentaire (donc aussi d'intersection).

code. *n.* Ensemble de symboles et de règles d'utilisation de ces symboles pour représenter des informations de façon conventionnelle.

codimension. *n.* Soit E un espace vectoriel. On dit qu'un sous-espace F est de codimension finie s'il admet un supplémentaire G dans E qui soit de dimension finie. Et l'on pose :
$$\text{codim } F = \dim G.$$

col. *n.* Voir *minimax.*

compact. *adj.* Un espace topologique E est dit compact si, de tout recouvrement ouvert de E, on peut extraire un recouvrement fini; il est dit localement compact s'il est séparé et si chacun de ses points possède au moins un voisinage compact.

compilateur. *n.* Programme destiné à traduire en langage machine binaire un programme écrit dans un langage évolué.

complément à deux. Mode de représentation des nombres négatifs. Dans une représentation à n bits, le nombre — N (N > 0) est représenté par $2^n - N$.

complémentaire (d'une partie dans un ensemble). Soit A une partie de l'ensemble E; l'ensemble des éléments de E qui n'appartiennent pas à A est le complémentaire de A dans E, noté $\complement_E^A$ ou E — A.

complet. *adj.* Un espace métrique E est dit complet si toute suite de Cauchy de points de E admet une limite dans l'espace E.

complexe. *adj.* **(nombre).** Se dit d'un nombre du type $a + bi$ où a et b sont des nombres réels et i une racine du polynôme $x^2 + 1$; a est la partie réelle, b la partie imaginaire.

compteur ordinal. Registre de l'unité de contrôle contenant l'adresse de la prochaine instruction à exécuter pendant le déroulement d'un programme. Son contenu est mis à jour à chaque instruction.

conforme. *adj.* **(transformation).** Transformation géométrique qui conserve les angles. Toute fonction d'une variable complexe, holomorphe, définit une transformation conforme.

congru. *adj.* Si $\mathcal{R}$ est la relation d'équivalence définie sur $\mathbb{Z}$ par $a \mathcal{R} b$ si et seulement si $a - b$ est un multiple de k (k entier non nul), on dit que a est congru à b modulo k et on note : $a \equiv b\ (k)$. *Ex. :* $5 \equiv 23\ (3)$.

congruence. *n.* Relation d'équivalence particulière définie sur $\mathbb{Z}$ (voir *congru*).

conique. *n.* Les coniques sont les courbes obtenues par intersection d'un cône et d'un plan. En coordonnées cartésiennes, ces courbes sont le lieu des points solutions des équations du second degré à deux inconnues.

continuité. *n.* Une application f d'un espace topologique E dans un espace topologique F est continue en un point $a \in E$ si $f(x)$ admet $f(a)$ pour limite lorsque x tend vers a.

contractante. *adj.* **(application).** Soit f une application d'un espace métrique (E, d) dans un autre (F, d') et k un nombre strictement compris entre 0 et 1 $(0 < k < 1)$. On dit que f est contractante si :
$$\forall x,\ y \in E,\ d'(f(x), f(y)) \leqslant k d(x, y).$$

contraction. *n.* Opération qui consiste à égaler dans un monôme un indice de covariance et un indice de contravariance, en sous-entendant alors la sommation. Un tenseur mixte (m, p) devient alors un tenseur mixte $(m - 1, p - 1)$.

convention d'Einstein. Convention qui permet de sous-entendre une sommation dans un monôme lorsque celui-ci comporte un indice répété deux fois, une fois en tant qu'indice de covariance, une autre comme indice de contravariance.

convergence. *n.* On dit qu'une suite (x_n) d'un espace métrique (E, d) converge vers un point $l \in E$ (ou admet l pour limite dans E) si pour tout $\varepsilon > 0$, il existe un entier p tel que, dès que n est un entier supérieur à p, on ait $d(l, x_n) < \varepsilon$.

convergence simple. Soit une suite f_n de fonctions à valeurs dans un espace métrique F; on dira que la suite f_n converge simplement vers f si :
$$\forall x \in E,\ \ \forall \varepsilon > 0\ \ \exists m \in \mathbb{N}$$
$$\forall n \geqslant m : d(f_n(x), f(x)) \leqslant \varepsilon.$$
L'entier m ainsi déterminé dépend de ε et de x.

convergence uniforme. Mode de convergence d'une suite de fonctions, tel que la distance entre un élément f_n de la suite et la limite devienne arbitrairement petite, mais indépendamment de la variable; pour tout $\varepsilon > 0$, il existe m tel que, dès que $n \geqslant m$, alors, et pour tout x, on a : $d(f_n(x), f(x)) \leqslant \varepsilon$.

convexe. *adj.* **(ensemble).** Une partie A d'un espace vectoriel E est dite convexe si, toutes les fois que deux points a, b appartiennent à A, tout le segment $[\vec{a}, \vec{b}]$ est contenu dans A : $\forall \vec{a}, \vec{b} \in A$, $\forall \lambda \in [0,1] : \lambda \vec{a} + (1 - \lambda) \vec{b} \in A$.

convexe. *adj.* **(figure).** Se dit d'une figure géométrique telle que si deux points lui appartiennent, alors le segment qui joint ces points lui appartient.

convolution. *n.* Loi de composition définie entre fonctions localement intégrables sur un ensemble X par :
$$(f, g) \mapsto f * g,$$
où $$f * g(x) = \int_X f(x - t)\, g(t)\, dt.$$

coordonnées. *n.* 1° *Coordonnées bipolaires.* Les coordonnées bipolaires d'un point P d'un plan sont les deux distances de P à deux points fixes appelés « pôles » ou « foyers ». 2° *Coordonnées cartésiennes.* Un système de référence cartésien étant fixé dans le plan, les coordonnées cartésiennes d'un point P du plan sont les mesures des deux segments ayant pour extrémités d'une part l'origine, d'autre part le point d'intersection avec l'un des axes de la droite parallèle à l'autre axe et passant par P. 3° *Coordonnées elliptiques.* Les coordonnées elliptiques d'un point P du plan dans lequel on a fixé deux foyers sont respectivement la somme et la différence des coordonnées bipolaires de P par rapport à ces deux foyers. 4° *Coordonnées polaires.* Si dans le plan est fixée une demi-droite orientée dont l'origine est le point O, munie d'un vecteur unitaire $\vec{i}$, les coordonnées polaires d'un point P du plan sont la distance $|\overrightarrow{OP}|$ et l'angle $(\vec{i}, \overrightarrow{OP})$.

corps. *n.* Un anneau sera appelé corps si ses éléments distincts du zéro de l'addition forment un groupe pour la multiplication.

corrélation (coefficient de). Mesure du degré de liaison entre deux variables aléatoires X et Y. Il existe plusieurs définitions des coefficients de corrélation (de Pearson, etc.); mais, sauf indication contraire, il s'agit du coefficient ρ défini pour tout couple X et Y admettant des espérances mathématiques E (X) et E (Y) et des écarts types σ (X) et σ (Y), tels que :
$$\rho = \frac{E\,[(X - E(X))\ (Y - E(Y))]}{\sigma(X)\, \sigma(Y)}.$$

corrélation (fonction de). La fonction de corrélation $C(\tau)$ est la fonction de τ formée par la covariance $E(y(t)\, y(t + \tau))$ pour une fonction aléatoire stationnaire $y(t)$.

corrélées. *adj.* **(variables aléatoires).** Se dit de deux variables aléatoires X et Y telles que le coefficient de corrélation entre X et Y soit différent de zéro. A distinguer de variables aléatoires *liées.*

correspondance. *n.* Soit E et F deux ensembles; on appelle correspondance de E dans F toute application de $E \to \mathfrak{P}(F)$ où $\mathfrak{P}(F)$ est l'ensemble des parties de F. Une correspondance Φ est généralement notée $\Phi : E \Rightarrow F$.

correspondances (analyse des). Méthode particulière d'analyse factorielle.

covariance. *n.* La covariance entre deux variables aléatoires X et Y est l'espérance mathématique, quand elle existe, du produit XY.

cycle. *n.* Une chaîne p-dimensionnelle dont le bord est nul est elle-même un bord; on l'appelle p-cycle.

D

décimal. *adj.* **(système).** Qualifie le système de numération à base 10.

définition imprédicative. Cette notion est apparue avec les paradoxes de la théorie des ensembles. On dit qu'un ensemble est défini de façon imprédicative si la définition renvoie à la

totalité à laquelle l'ensemble appartient. En langage logique, une définition est imprédicative si elle définit un objet qui est une des valeurs prises par une variable liée occurente dans l'expression symbolique de la définition. Poincaré et Russell ont essayé, de différente façon, d'éviter les définitions imprédicatives.

dénombrable. *adj.* **(ensemble).** Se dit d'un ensemble équipotent à l'ensemble $\mathbb{N}$ des entiers naturels.

dense. *adj.* Une partie A d'un espace topologique E est dite dense dans E si sa fermeture est égale à E.

densité de probabilité. La densité de probabilité $f(x)$, définie sur l'intervalle de définition de la variable aléatoire continue X, $[a, b]$, $a \neq b$, est telle que :

$$Pr (x \leqslant X \leqslant x + dx) = f(x)\, dx \quad (x \in [a, b]).$$

densité de probabilité normale. La variable aléatoire X continue, définie sur $(-\infty, +\infty)$, est normale si la densité de probabilité $f(x)$ associée peut se mettre sous la forme :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-x_o)^2}{2\sigma^2}}$$

où $x_o \in R$ et $\sigma \in R_+$.

dérivée. *n.* Soit I un intervalle de $\mathbb{R}$ et f une application de I dans $\mathbb{R}$. Lorsque la limite du rapport :

$$\frac{f(x_o + h) - f(x_o)}{h}$$

existe pour h tendant vers 0 (avec $x_o \in$ I), on dit que f est dérivable en x_o. L'application qui, à tout $x \in$ I, fait correspondre la limite obtenue s'appelle application dérivée ou dérivée de f.

déterminant. *n.* Sur un espace vectoriel E, seule forme multilinéaire alternée qui prenne la valeur + 1 pour les vecteurs de la base canonique.

dialectique. *n.* Au sens où on l'a utilisé dans le texte, ce terme, qui a eu droit à la plus grande fortune, désigne le discours par lequel on essayait de donner aux mathématiques le statut qui leur revient dans le « monde intelligible ». On lit souvent que la dialectique trouve son origine dans les discussions antiques dont Zénon, disciple de Parménide, nous a laissé le souvenir dans les fameux paradoxes associés à son nom. Mais, déjà avec Aristote, le terme dialectique déchoit des hauteurs platoniciennes : on l'applique aux syllogismes dont les prémisses ne sont pas nécessairement vraies.

différence symétrique (de deux ensembles). C'est l'ensemble des éléments qui appartiennent à l'un seulement des deux ensembles, noté :

$$A \triangle B = \complement_{A \cup B} (A \cap B).$$

différentielle. *n.* Application linéaire continue l, approximant la différence des valeurs prises par une fonction f (définie sur un ouvert de $\mathbb{R}^n$) en deux points, donc telle que $\|f(x) - f(y) - l(x - y)\|$ soit négligeable devant $\|x - y\|$.

dimension. *n.* Toutes les bases d'un espace vectoriel ont le même nombre d'éléments; ce nombre est appelé dimension de l'espace vectoriel.

Dirac (distribution de). La densité de probabilité de Dirac $\delta(x - x_o)$ est la densité de probabilité associée à la variable aléatoire X continue sur $(-\infty, +\infty)$ telle que $Pr (X \neq x_o) = 0$. Elle est définie par la propriété fondamentale de la distribution de Dirac :

$$\int_{-\infty}^{+\infty} y(x)\, \delta(x - x_o)\, dx = y(x_o),$$

valable pour toute fonction $y(x)$ à valeurs dans $\mathbb{R}$.

discriminante. *adj.* **(analyse).** Qualifie la méthode statistique de traitement de données multidimensionnelles à but explicatif ou prévisionnel.

disjoints. *adj.* **(sous-ensembles).** Se dit de sous-ensembles dont l'intersection est vide.

distance. *n.* Une distance d sur un ensemble E est une application de $E \times E$ dans l'ensemble des nombres réels positifs ou nuls, telle que, quels que soient les points x, y et z appartenant à E, on ait :

$$d(x, y) = 0 \Leftrightarrow x = y;$$
$$d(x, y) = d(y, x);$$
$$d(x, y) \leqslant d(x, z) + d(z, y).$$

On dit encore métrique définie sur E.

distribution. *n.* Soit $\mathfrak{D}$ l'espace vectoriel des fonctions à valeurs réelles définies sur $\mathbb{R}^n$, indéfiniment différentiables et à support compact. On dit qu'une suite (φ_p) de $\mathfrak{D}$ converge vers O dans l'espace $\mathfrak{D}$ si les supports des fonctions φ_p sont contenus dans un même compact et si toutes les dérivées partielles successives convergent uniformément vers O sur $\mathbb{R}^n$. On appelle distribution tout élément T du dual topologique de l'espace $\mathfrak{D}$ muni de la topologie sus-indiquée.

diviseur. *n.* Dans un anneau, a est un diviseur de b s'il existe un élément k de l'anneau tel que $b = ka$. *Ex. :* dans $\mathbb{Z}$, 3 est un diviseur de 12.

dual. *n.* Ensemble des formes linéaires sur un espace vectoriel E; pour les opérations d'addition et de produit par un scalaire il possède une structure d'espace vectoriel. On le note E^*.

dual topologique. Soit E un espace vectoriel topologique. On appelle dual topologique de E l'espace E' des formes linéaires continues sur E.

duale. *adj.* **(base).** Si $(e_i)_{i \in I}$ désigne une base d'un espace vectoriel E, les formes linéaires $(e^j)_{j \in I}$ telles que $e^j(e_i) = 0$ si $i \neq j$ et $+1$ si $i = j$ forment une base de E^*, dite base duale de celle de E.

E

écart type. Racine positive de la variance.

échantillon. *n.* L'échantillon $\{x_1 \ldots x_n\}$ de taille n est le résultat du tirage de la variable aléatoire à n dimensions $\{X_1 \ldots X_n\}$, telle que les variables aléatoires X_i soient mutuellement indépendantes et soient régies par la même loi de probabilité. Le tirage répété n fois d'une variable aléatoire X de loi $F(x)$ est donc une méthode possible pour la constitution d'un échantillon de taille n.

éditeur de liens. Voir *chargeur*.

effet Condorcet. Voici le texte de Condorcet où apparaît ce concept : « Supposons, en effet, toujours trois candidats, et que les électeurs soient au nombre de soixante; qu'il y ait vingt-trois voix pour l'ordre Pierre, Paul, Jacques; aucune pour l'ordre Pierre, Jacques, Paul; deux pour l'ordre Paul, Pierre, Jacques; dix-sept pour l'ordre Paul, Jacques, Pierre; dix pour l'ordre Jacques, Pierre, Paul et huit pour l'ordre Jacques, Paul, Pierre; (...). Les trois propositions adoptées par la pluralité seraient donc :
— Pierre est préférable à Paul;
— Jacques est préférable à Pierre;
— Paul est préférable à Jacques.
Et il est évident que ces trois propositions ne peuvent être vraies en même temps, puisque (...) de deux quelconques, admises ensemble, résulte nécessairement une conséquence contradictoire avec la troisième... (si bien que) l'élection donne un résultat absurde, ou plutôt elle n'en donne aucun, tandis que, suivant la méthode ordinaire, Pierre ayant vingt-trois, Paul dix-neuf et Jacques dix-huit voix, l'élection décide en faveur de Pierre. »

Élée, éléates. Élée est une ville du sud de l'Italie, sur les bords de la mer Tyrrhénienne. Colonie phocéenne, elle est la patrie des philosophes « éléates », Zénon et Parménide en particulier.

élément neutre. Un élément e d'un ensemble E muni d'une loi de composition interne (notée T) est un élément neutre pour la loi T si, pour tout élément x de E, on a les égalités :

$$x \,\mathrm{T}\, e = e \,\mathrm{T}\, x = x.$$

élément symétrique. Si E est un ensemble muni d'une loi de composition interne (notée T) possédant un élément neutre e, on dira qu'un élément x' de E est symétrique d'un élément x de E si :

$$x \,\mathrm{T}\, x' = x' \,\mathrm{T}\, x = e.$$

ellipse. *n.* C'est une conique dont l'équation peut se mettre sous la forme :

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$$

ellipsoïde. *n.* C'est une quadrique qui peut se représenter par l'équation :

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1.$$

endomorphisme. *n.* Application linéaire d'un espace vectoriel dans lui-même.

ensemble. *n.* Mot le plus général pour désigner un « groupement » d'objets; mais tout groupement d'objets n'est pas forcément un ensemble (voir *axiomes de la théorie des ensembles*).

ensemble des parties d'un ensemble. C'est l'ensemble dont les éléments sont les parties de l'ensemble E considéré; il est noté $\mathfrak{P}$ (E) et comprend les parties propres, la partie vide et la partie pleine.

ensemble-quotient. Ensemble des classes d'équivalence d'un ensemble E muni d'une relation d'équivalence $\mathfrak{R}$; on le note $E/\mathfrak{R}$. *Ex. :* si $E = \mathbb{Z}$ et si $\mathfrak{R}$ est la congruence modulo 2, $E/\mathfrak{R}$ se réduit à deux classes : celle des nombres pairs et celle des nombres impairs.

entier relatif (nombre). Nombre égal à un nombre entier ou à son opposé.

entrées-sorties. Toutes les opérations d'échange d'information entre l'unité centrale et les organes périphériques en informatique.

épistémologie. *n.* Ce terme est introduit en France au début du XX[e] siècle par Couturat dans le lexique qu'il joint à la traduction par Cadenat de l'ouvrage de Russell : *An Essay of the Foundations of Geometry* (1897). Couturat, inspiré par la philosophie kantienne, définit l'épistémologie comme la « théorie de la connaissance appuyée sur l'étude critique des Sciences ». Bientôt, Lalande et Robert consignent à peu près la même définition dans leurs dictionnaires respectifs, et Meyerson consacre l'usage définitif du terme en 1907 (dans « Identité et Réalité »).
Après les *Fondements de la Géométrie* de Hilbert, les études épistémologiques se sont émancipées de toute autorité philosophique. L'idée même de « théorie de la connaissance » est critiquée, et l'esprit de système, banni du corps des sciences, ne peut plus sévir dans les études qui prennent les sciences pour objet (voir *Critique*, n° 308, janvier 1973, p. 53-66).

équation. *n.* Soit f et g deux applications d'un ensemble E dans un ensemble F. La relation :

$$f(x) = g(x)$$

s'appelle *équation* et tout élément x de E pour lequel la relation est vraie s'appelle solution de l'équation. Résoudre l'équation, c'est déterminer l'ensemble K de toutes les solutions :

$$K = \{x \mid f(x) = g(x)\}.$$

équation différentielle. Soit E un espace vectoriel normé sur $\mathbb{R}$ et f une application continue sur $\mathbb{R} \times E$ à valeurs dans E. L'équation $\frac{dy}{dt} = f(t, y)$ s'appelle équation différentielle du premier ordre. Une solution de cette équation est une application φ dérivable sur un intervalle I de $\mathbb{R}$ et telle qu'en tout point x de I l'on ait :

$$\varphi'(x) = f(x, \varphi(x)).$$

Plus généralement, l'équation :

$$\frac{d^n y}{dt^n} = f\left(t, y, \frac{dy}{dt}, \ldots, \frac{d^{n-1} y}{dt^{n-1}}\right)$$

s'appelle équation différentielle d'ordre n.

équation linéaire. Soit E et F deux espaces vectoriels; on appelle équation linéaire toute équation de la forme : $f(\vec{x}) = \vec{0}_F$ où f est une application linéaire de E dans F.

équidécomposable. *adj.* Deux figures sont dites équidécomposables si on peut les décomposer en un même nombre de triangles respectivement égaux.

équipotent. *adj.* Deux ensembles sont équipotents s'ils peuvent être mis en bijection.

équivalence (classe d'). Sous-ensemble constitué par des éléments équivalents.

équivalence (relation d'). Relation binaire réflexive, symétrique et transitive.

ergodique. *adj.* **(chaîne).** Chaîne markovienne telle qu'il n'existe pas d'états qui ne puissent être atteints à partir d'un état initial quelconque. Si P est la matrice stochastique associée, de rang $m \times n$, on a donc :

$$\forall i,\ 1 \leqslant i \leqslant m;$$
$$\forall j,\ 1 \leqslant j \leqslant n;$$
$$\exists n \in \mathbb{N}^{+}_{*}\ (P^n)_{ij} \neq 0.$$

espace de Banach. Espace vectoriel normé complet.

espace de Hilbert. On appelle espace de Hilbert un espace vectoriel E muni d'une forme bilinéaire symétrique dont la forme quadratique associée est définie positive; cette forme quadratique définit une norme par laquelle l'espace E est complet.

espace localement convexe. Un espace vectoriel topologique E est dit localement convexe si tout point admet un système fondamental de voisinages convexes.

espace probabilisable. $\mathcal{E}$ est un espace probabilisable s'il est un espace d'événements possibles tel qu'à tout élément e de $\mathcal{E}$ et à tout élément d'une tribu associée à $\mathcal{E}$ on peut associer une probabilité intrinsèque de réalisation.

espace réflexif. On dira qu'un espace normé E est réflexif s'il est identique à son bidual E''.

espace vectoriel normé. Soit E un espace vectoriel sur $\mathbb{R}$. On appelle alors norme sur E toute fonction, notée $\vec{x} \rightarrow \|\vec{x}\|$, possédant les trois propriétés suivantes :

1° $\|\vec{x}\| > 0$ pour $\vec{x} \neq \vec{0}$; $\|\vec{0}\| = 0$.
2° $\|\lambda.\vec{x}\| = |\lambda|\ \|\vec{x}\|$, $\lambda \in \mathbb{R}$.
3° $\|\vec{x} + \vec{y}\| \leqslant \|\vec{x}\| + \|\vec{y}\|$.

Si E est muni d'une telle norme, on l'appelle espace vectoriel normé. Un espace vectoriel normé est un espace topologique.

espace vectoriel semi-normé. Un espace vectoriel est dit semi-normé s'il est muni d'une famille de semi-normes $(P_i)_{i} \in I$ ayant la propriété suivante : quel que soit l'ensemble fini $J \in I$, il existe k tel que P_k majore tous les P_j, $j \in J$.

espace vectoriel topologique. On appelle espace vectoriel topologique un ensemble E muni d'une structure espace vectoriel sur $\mathbb{R}$ et, d'autre part, d'une topologie compatible avec la structure vectorielle, c'est-à-dire telle que l'addition $\vec{x}, \vec{y} \rightarrow \vec{x} + \vec{y}$ est *continue* de $E \otimes E \rightarrow E$ et que la multiplication $\lambda, \vec{x} \rightarrow \lambda\vec{x}$ est également *continue* de $\mathbb{R} \otimes E \rightarrow E$.

espaces de Sobolev. Soit Ω un ouvert de $\mathbb{R}^n$; on appelle espace de Sobolev d'ordre 1 l'espace défini par :

$$H^1(\Omega) = \left\{ v | v \in L^2(\Omega), \frac{\partial v}{\partial x_1} \in L^2(\Omega), \ldots \ldots \frac{\partial v}{\partial x_n} \in L^2(\Omega) \right\}$$

où $\frac{\partial v}{\partial x_i}$ désigne la dérivée au sens des distributions par rapport à x_i.
Plus généralement, on définit des espaces de Sobolev d'ordre n, en faisant intervenir les dérivées partielles successives d'ordre $\leqslant n$.

espérance mathématique. Soit X une variable aléatoire. L'espérance mathématique E (X) de X est :

$$E(X) = \sum_i x_i F(x_i)$$

si X est une variable aléatoire discrète et $F(x_i)$ la loi de probabilité associée; ou bien elle est :

$$E(X) = \int_a^b x f(x)\, dx$$

si X est une variable aléatoire continue, définie sur l'intervalle $[a, b]$, et si $f(x)$ est la densité de probabilité associée.

estimateur. *n.* Soit $\{X_1 \ldots X_n\}$ une suite de variables aléatoires mutuellement indépendantes et régies par la même loi de probabilité $F_\alpha(x)$, cette dernière dépendant d'un paramètre α. On appelle estimateur $\widehat{\alpha_n}$ de α toute variable aléatoire $\widehat{\alpha_n} = f(X_1 \ldots X_n)$ telle que, si $\{x_1 \ldots x_n\}$ est un échantillon de taille n résultat du tirage de $\{X_1 \ldots X_n\}$, $\alpha^*_n = f(x_1 \ldots x_n)$ puisse être considéré comme une valeur approchée de α.

estimateur asymptotiquement sans biais. Estimateur $\widehat{\alpha_n}$ d'un paramètre α tel que :

$$\lim_{n \to \infty} E(\widehat{\alpha_n}) = \alpha.$$

estimateur sans biais. Estimateur $\widehat{\alpha_n}$ d'un paramètre α tel que :

$$\forall_n\ E(\widehat{\alpha_n}) = \alpha.$$

explicative. *adj.* **(variable).** Variable intervenant dans les problèmes de régression; appelée aussi *variable exogène.*

expliquée. *adj.* **(variable).** Variable intervenant dans les problèmes de régression; appelée aussi *variable endogène.*

F

face. *n.* Dans un p-simplexe S, on appelle face chacun des $(p - 1)$-simplexes formés par $p - 1$ sommets de S. Un p-simplexe possède donc $p - 1$ faces.

facteur. *n.* Nouvelle variable résultant d'une analyse factorielle.

factoriel. *adj.* **(axe).** Se dit d'un axe associé à un vecteur propre d'une matrice, étudiée par une méthode d'analyse factorielle.

factorielle. *adj.* **(analyse).** Qualifie une méthode statistique de traitement de données multidimensionnelles. On distingue l'analyse en facteurs communs et facteurs spécifiques, l'analyse en composantes principales et l'analyse des correspondances.

fermé. *adj.* Complémentaire d'un sous-ensemble ouvert dans un espace topologique.

fermeture. *n.* La fermeture (ou adhérence) d'une partie A d'un espace topologique E est le plus petit fermé de E qui contienne A.

finitisme de Hilbert. On appelle ainsi l'exigence de Hilbert, qui ne voulait admettre dans la théorie de la preuve (ou métamathématique) que des arguments finitistes, c'est-à-dire :
1° qui ne considèrent que des collections *finies* d'objets ou de fonctions *bien définies,* et donc dont on peut calculer univoquement la valeur;
2° qui n'affirment l'existence d'aucun objet sans indiquer en même temps le moyen de le construire;
3° qui ne parlent jamais de l'ensemble de *tous* les objets x d'une collection *infinie* (quand on dit qu'un théorème vaut pour tous ces x, cela veut dire seulement que pour chaque x, en particulier, on peut répéter le raisonnement qui l'établit en général).

flip flop. *n.* Circuit électronique à deux états permettant de mémoriser une information logique; appelé aussi *bistable.*

foncteur. *n.* Un foncteur d'une catégorie C dans une catégorie C' est une application F de l'ensemble C dans l'ensemble C' telle que :
1° $F[\alpha(f)] = \alpha[F(f)]$ et $F[\beta(f)] = \beta[F(f)]$;
2° si $g \circ f$ est défini dans C, alors :

$$F(g \circ f) = F(g) \circ F(f).$$

fonction. *n.* Relation telle que tout élément de l'ensemble de départ ait au plus une image. Si l'on réduit l'ensemble de départ à l'ensemble de définition de la fonction, la fonction est alors une application. On utilise plutôt le terme fonction lorsque l'ensemble d'arrivée est $\mathbb{R}$ (fonction numérique) ou $\mathbb{C}$ (fonction complexe).

fonction convexe, concave. Soit f une application d'un espace vectoriel E dans $\mathbb{R}$; on dit que f est une application convexe si $\forall \vec{x}, \vec{y} \in E$, $\forall \lambda \in [0, 1]$, on a l'inégalité :

$$f(\lambda\vec{x} + (1 - \lambda)\vec{y}) \leqslant \lambda f(\vec{x}) + (1 - \lambda) f(\vec{y}).$$

On dira qu'elle est concave si $-f$ est convexe.

fonction logique. Fonction ne pouvant prendre que deux valeurs, vrai ou faux, selon les valeurs des variables.

forme bilinéaire définie positive. Une forme bilinéaire définie sur $E \times E$ (E espace vectoriel) est dite définie positive si :
1° $B(\vec{x}, \vec{x}) \geqslant 0 \quad \forall \vec{x} \in E$;
2° $B(\vec{x}, \vec{x}) = 0 \Rightarrow \vec{x} = \vec{0}$.

frontière. *n.* On appelle frontière d'une partie A d'un espace topologique E l'intersection de l'adhérence de A et du complémentaire de l'intérieur de A :

$$Fr(A) = \bar{A} \cap \complement_E \mathring{A}$$

G

galoisienne. *adj.* **(extension).** Se dit d'une extension algébrique simple L d'un corps K telle que, si un polynôme de degré n, irréductible dans K, y admet une racine, alors il en possède n et s'y décompose donc en un produit de n facteurs irréductibles du premier degré.

gaussienne. *adj.* **(variable aléatoire).** Une variable aléatoire X est dite gaussienne si elle est définie sur $(-\infty, +\infty)$ et si la densité de probabilité associée est normale.

génératrice. *adj.* **(partie).** Se dit d'une partie G d'un espace vectoriel E, telle que tout élément de E puisse s'obtenir comme combinaison linéaire d'éléments de G.

gradient. *n.* Vecteur à n composantes et dont les coordonnées sont les dérivées partielles du premier ordre d'une fonction réelle définie sur un ouvert de $\mathbb{R}^n$.

groupe. *n.* Un groupe est un ensemble G muni d'une loi de composition interne partout définie, et qui possède les propriétés suivantes :
1° elle est associative;
2° elle possède un élément neutre;
3° tout élément de G admet un symétrique unique.

groupe commutatif. Un groupe est commutatif (ou abélien) si sa loi de composition interne est commutative.

groupe de substitutions. La théorie des groupes de substitutions ou, comme on dit aussi, de permutations renvoie à la théorie de Galois dont on trouvera un bref résumé dans le chapitre *Histoire des Mathématiques,* au paragraphe consacré à l'algèbre au XIXe siècle.

groupe quotient. Si H est un sous-groupe distingué d'un groupe G, l'ensemble quotient $G/\mathcal{R}$, où $\mathcal{R}$ est la relation d'équivalence $x\mathcal{R}y$ si et seulement si $x^{-1}y \in H$, est un groupe pour la loi $\bar{x}\bar{y} = \overline{xy}$ ($\bar{x}$ désigne un élément de $G/\mathcal{R}$, x un élément de G et xy le composé dans G de x et de y). Ce groupe est appelé groupe quotient de G par le sous-groupe distingué H.

H

hasard. *n.* Antécédent de tout événement ou de tout phénomène dont la manifestation ne peut pas être prédite à l'avance, soit par ignorance des causes présidant à sa réalisation, soit par indéterminisme fondamental.

hiérarchique. *adj.* **(classification).** Se dit de la méthode statistique consistant à déterminer des classes sur un ensemble de données multidimensionnelles par un procédé itératif.

holomorphe. *adj.* Qualifie une fonction dérivable de variable complexe. Toute fonction d'une variable complexe, holomorphe, est analytique et *vice versa.*

homéomorphe. *adj.* Deux espaces topologiques sont homéomorphes lorsqu'il existe un homéomorphisme de l'un sur l'autre. De tels espaces ont les mêmes propriétés topologiques.

homéomorphisme. *n.* Bijection continue d'un espace topologique sur un autre espace topologique, telle que sa réciproque soit aussi une application continue.

homologie (groupe d'). Ensemble des classes d'équivalence de p-cycles pour la relation : $A \sim B$, s'il existe une $(p+1)$-chaîne C, telle que A-B soit le bord de C.

homologue. *adj.* Deux segments, deux angles appartenant à deux figures géométriques sont dits homologues s'il existe, entre les deux figures, une correspondance biunivoque qui les fait correspondre.

homomorphisme d'anneaux. Une application f d'un anneau A dans un anneau A' est un homomorphisme d'anneaux si :
— $f(x+y) = f(x) + f(y)$;
— $f(xy) = f(x) f(y)$ pour tous les éléments x, y de A.

homomorphisme de groupes. Une application f d'un groupe G dans un groupe G' est un homomorphisme de groupes si :

$$f(xy) = f(x) f(y)$$

pour tout x, tout y appartenant à G (les lois internes de G et G' étant notées multiplicativement).

homotope. *adj.* Deux applications continues (ou *arcs*) f et g de l'intervalle [0, 1] dans un espace E sont dites homotopes s'il existe une « déformation continue de f sur g conservant les extrémités », $H : [0, 1] \times [0, 1] \to E$ et telle que :

$$\begin{cases} H(t, 0) = f(t) \\ H(t, 1) = g(t). \end{cases}$$

hyperbole. *n.* Une hyperbole est une conique dont l'équation peut se mettre sous la forme :

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1.$$

hyperboloïde à deux nappes. L'hyperboloïde à deux nappes est une quadrique qui peut se représenter par l'équation :

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} - \frac{z^2}{c^2} = 1.$$

hyperboloïde à une nappe. L'hyperboloïde à une nappe est une quadrique qui peut se représenter par l'équation :

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = 1.$$

hyperplan. *n.* Soit un espace vectoriel E. On appelle hyperplan de E tout sous-espace de E de codimension 1.

hypothético-déductif. *adj.* Une science est dite « hypothético-déductive » quand elle peut s'organiser en un ensemble d'énoncés dont un petit nombre, admis au titre d'hypothèses, permet, moyennant l'usage de règles déterminées, d'obtenir par des déductions plus ou moins longues et complexes les autres énoncés. En ce sens, toute science hypothético-déductive doit son caractère à l'application de la méthode axiomatique (voir *axiomatisation*).

I

idéal. *n.* Un idéal à gauche d'un anneau A est une partie I de A vérifiant :
— I est un sous-groupe du groupe additif de A;
— pour tout élément x de A, $xI = I$ (pour un idéal à droite, la seconde propriété devient $Ix = I$).

idempotent. *adj.* Dans un ensemble E muni d'une loi T, un élément a est idempotent pour la loi T si $aTa = a$.

image. *n.* Si f est une application linéaire d'un espace vectoriel E dans un espace vectoriel F, l'ensemble $f(E)$ est appelé image de f et se note $Im(f)$.

imaginaire. *adj.* **(nombre).** Se dit d'un nombre complexe dont la partie réelle est nulle. *Ex. :*

$$i\sqrt{2}, \quad -3i.$$

inclusion. *n.* Relation entre deux ensembles, qui exprime que tout élément de l'un des ensembles (A) est aussi élément de l'autre ensemble (B) ; se note $\subset$. *Ex. :*

$$A \subset B \Leftrightarrow [x \in A \Rightarrow x \in B].$$

indépendantes. *adj.* **(variables aléatoires).** Se dit d'un ensemble de variables aléatoires $\{X_k\}_{k=1, N}$ définies sur des espaces d'événements possibles $\mathcal{E}_1, \ldots \mathcal{E}_N$ telles que pour tous $i, j \leqslant N$:

$$\forall a_i \in \mathcal{E}_i \text{ et } a_j \in \mathcal{E}_j$$
$$Pr(a_i \supset a_j) = Pr(a_i) \, Pr(a_j).$$

indexage. *n.* Mode d'adressage consistant à ajouter à l'adresse spécifiée dans une instruction le contenu du registre d'index. On peut ainsi, dans une boucle, atteindre des adresses successives avec la même instruction en incrémentant ce contenu à chaque itération.

indice. *n.* La valeur (entière) de l'intégrale :

$$\frac{1}{2\pi i} \int_\gamma \frac{dz}{z-a}$$

représente le nombre algébrique de fois où la courbe γ tourne autour du point a, et s'appelle l'indice de a par rapport à la courbe γ.

indirect. *adj.* **(adressage.)** Se dit de la méthode d'accès à une information consistant à indiquer dans une instruction non pas l'adresse de l'information cherchée, mais l'adresse d'un mot contenant l'adresse de l'information.

inductif. *adj.* **(ensemble).** Un ensemble ordonné E est inductif si toute partie totalement ordonnée de E admet un majorant.

infini. *n.* ou *adj.* Dès sa naissance comme science déductive, la mathématique a rencontré la notion d'infini, sous diverses formes; on distingue traditionnellement (depuis Aristote) l'*infini actuel* et l'*infini potentiel*. Jusqu'au XIXe siècle, seul ce dernier avait un statut mathématique : il désigne la possibilité de poursuivre indéfiniment un processus opératoire. Aristote pensait que les mathématiciens n'ont pas besoin de recourir à l'infini mais seulement à une grandeur finie aussi grande que nécessaire. Par infini « actuel » ou « en acte », on désigne un infini donné, un « être ». C'est pourquoi les premiers « actualistes » ont été des théologiens, critiques à l'égard de la doctrine d'Aristote qui ne permettait ni au monde ni à Dieu d'être infinis. Mathématiquement, les totalités infinies actuellement données ont reçu droit de cité avec la création d'une arithmétique de l'infini.

instruction. *n.* Ordre d'exécution d'une opération élémentaire réalisable par le calculateur.

intégrale. *adj.* **(courbe).** Famille de solutions d'une équation différentielle d'ordre n, définies implicitement et dépendant de n paramètres.

intégrale de Lebesgue. Généralisation de la notion d'intégrale au cas de certaines fonctions pour lesquelles on ne peut définir une intégrale au sens de Riemann.

intégrale de Riemann-Stieltjes. Forme linéaire continue pour la topologie de la convergence uniforme, sur l'espace de Banach des fonctions numériques définies et continues sur un intervalle de $\mathbb{R}$.

intérieur. *n.* L'intérieur d'une partie A d'un espace topologique E est le plus grand ouvert contenu dans A.

intérieur relatif. Soit E un espace vectoriel topologique et A une partie de E; on appelle intérieur relatif de A (et l'on note $r_i A$) l'intérieur de A considéré comme sous-ensemble de $\mathcal{A}$, où $\mathcal{A}$ désigne la plus petite variété affine contenant A — $\mathcal{A}$ étant muni de la topologie induite par celle de E.

intersection. *n.* Pour deux ensembles, désigne l'ensemble de leurs éléments communs; désigne également l' « opération » qui associe à deux ensembles leur intersection, notée $\cap$. *Ex. :*

$$x \in A \cap B \Leftrightarrow [x \in A \textbf{ et } x \in B].$$

irrationnel. *adj.* **(nombre).** Se dit d'un nombre qui ne peut pas être représenté par le quotient de deux nombres entiers. *Ex. :* $\sqrt{2}$, π.

irréductible. *adj.* **(fraction).** Qualifie une fraction constituée par le quotient de deux nombres entiers premiers entre eux.

isométrie. *n.* Correspondance biunivoque entre deux figures géométriques, telle que les segments et les angles déterminés par les points qui se correspondent soient égaux. Deux figures entre lesquelles il existe une isométrie sont dites isométriques ou égales.

isomorphisme. *n.* Un isomorphisme est un homomorphisme (de groupes, d'anneaux, de corps) bijectif.

L

langage machine. Code binaire représentant les instructions utilisées par le calculateur.

langage symbolique. Langage permettant d'écrire des programmes au moyen de symboles sans avoir à connaître le langage machine.

libre. *adj.* **(partie).** Se dit de la partie d'un espace vectoriel dont aucun élément n'est combinaison linéaire des autres.

liées. *adj.* **(variables aléatoires).** Variables aléatoires qui ne sont pas indépendantes.

limite. *n.* Une application f d'un espace topologique E dans un autre espace topologique F tend vers la limite l quand x tend vers a par valeurs dans A ($A \subset E$), si pour tout voisinage W de l dans F il existe un voisinage V de a dans E tel que :

$$f(V \cap A) \subset W.$$

linéaire. *adj.* **(application).** Une application f d'un espace vectoriel E dans un espace vectoriel F (tous deux sur le même corps K) est dite linéaire si la propriété $f(aX + bY) = af(X) + bf(Y)$ est vérifiée pour tous les éléments a et b de K et tous les éléments X et Y de E.

linéaire. *adj.* **(forme).** Application linéaire d'un espace vectoriel dans son corps de base.

locale. *adj.* **(notion** ou **propriété).** Notion ou propriété topologique qu'il suffit de vérifier dans un voisinage de chaque point, et non de façon globale.

logique du premier ordre et **logique du second ordre.** Une *logique du premier ordre* se caractérise par le fait qu'elle n'admet comme seules variables que les variables d'individus, tandis qu'*au second ordre* sont admises des variables de prédicat : on peut alors appliquer les quantificateurs non plus seulement à des variables d'individus mais à des variables de sous-ensembles, et considérer par exemple une relation $\mathcal{R}$ (qui est un certain sous-ensemble d'un produit cartésien d'ensembles donnés) ou une suite $(x_n)_{n \in \mathbb{N}}$ (qui, en tant qu'application d'un ensemble E dans $\mathbb{N}$, est encore un certain sous-ensemble du produit $E \times \mathbb{N}$) comme des variables quantifiables.

loi de composition externe. On appelle loi de composition externe entre éléments d'un ensemble Ω (ensemble des opérateurs) et éléments d'un ensemble E une application de $\Omega \times E$ dans E.

loi de composition interne. Une loi de composition interne dans un ensemble E est une application d'une partie A de $E \times E$ dans E. 1° *Loi associative.* Une loi de composition interne (notée T) partout définie sur un ensemble E est associative si, pour tous les éléments x, y, z de E, on a : $(x \top y) \top z = x \top (y \top z)$. 2° *Loi commutative.* Une loi de composition interne (notée T) partout définie sur un ensemble E est commutative si, pour tous les éléments x, y de E, on a l'égalité :

$$x \top y = y \top x.$$

loi de « participation mystique ». Cette loi est particulièrement efficiente dans les danses et les cérémonies rituelles des « primitifs ». La raison profonde de ces cérémonies, pour ceux qui les célèbrent comme pour ceux qui y assistent, est, nous dit Lucien Lévy-Bruhl (*le Surnaturel et la Nature dans la Mentalité primitive*, P.U.F., Paris, 1963, p. 129), la « communion, la fusion mystique qui les identifie, suivant les cas, avec l'ancêtre mythique ou totémique,

homme-animal ou homme-plante, ou avec les "génies" des espèces animales et végétales, ou avec les ancêtres et les morts du groupe... »

loi des trois états. A. Comte la définit de la façon suivante : « En étudiant ainsi le développement total de l'intelligence humaine dans ses diverses sphères d'activité, depuis son premier essor le plus simple jusqu'à nos jours, je crois avoir découvert une grande loi fondamentale, à laquelle il est assujetti par une nécessité invariable, et qui me semble être solidement établie, soit sur les preuves rationnelles fournies par la connaissance de notre organisation, soit sur les vérifications historiques résultant d'un examen attentif du passé. Cette loi consiste en ce que chacune de nos conceptions principales, chaque branche de nos connaissances passe successivement par trois états théoriques différents : l'état théologique, ou fictif; l'état métaphysique, ou abstrait; l'état scientifique, ou positif. En d'autres termes, l'esprit humain, par sa nature, emploie successivement dans chacune de ses recherches trois méthodes de philosopher, dont le caractère est essentiellement différent et même radicalement opposé : d'abord la méthode théologique, ensuite la méthode métaphysique et enfin la méthode positive. De là, trois sortes de philosophies, ou de systèmes généraux de conceptions sur l'ensemble des phénomènes, qui s'excluent mutuellement : la première est le point de départ nécessaire de l'intelligence humaine, la troisième son état fixe et définitif; la seconde est uniquement destinée à servir de transition. » (Première leçon du *Cours de Philosophie positive*, éd. Hermann, 1975, Paris, p. 21.)

M

maïeutique. *n.* Il s'agit de l'art dont Socrate s'attribue (dans le *Théétète* de Platon) la spécialité et qui consiste à « accoucher » les esprits des pensées qu'ils contiennent à leur insu.

majorant. *n.* Dans un ensemble ordonné E, un majorant d'un sous-ensemble A est un élément M de E supérieur ou égal à tous les éléments de A.

markovienne. *adj.* **(chaîne).** Qualifie une chaîne de variables aléatoires $\{x_1, \dots x_k, \dots x_n\}$ telle que x_k $(2 < k \leqslant n)$ soit indépendante de toutes les variables x_i, $i \neq k$, à l'exception de x_{k-1}.

matrice. *n.* Mode de représentation d'une application linéaire, par lequel on range dans un tableau une succession de vecteurs-colonne, images des vecteurs d'une base de l'espace de départ exprimés par leurs composantes dans une base de l'espace d'arrivée.

matrice jacobienne. Soit f une application définie sur un ouvert U de $\mathbb{R}^p$ à valeurs dans $\mathbb{R}^n$, différentiable en un point $\vec{a} = (a_1, \dots a_p)$ de U. La matrice b_{ij} associée à l'application linéaire Df $(\vec{a})$ est définie par la relation :

$$b_{ij} = \frac{\partial f_i}{\partial x_j}\,(\vec{a})$$

où $(f_i)_{1 \leqslant i \leqslant n}$ désigne la famille des composantes de f; une telle matrice est appelée matrice jacobienne associée à f au point $\vec{a}$.

matrice de jeu. Matrice attachée à un jeu à deux personnes U dont l'élément U_{ij} est l'utilité à laquelle conduit l'application par le joueur A de la stratégie A_i et par le joueur B de la stratégie B_j.

maximal. *adj.* Un élément a d'un ensemble ordonné est maximal s'il n'admet pas de majorant strict, autrement dit si $x \geqslant a \Rightarrow x = a$.

médiane. *adj.* **(valeur).** La valeur médiane de la variable aléatoire X admettant la densité de probabilité $f(x)$ définie sur l'intervalle $[a, b]$, est la valeur $x_o \in [a, b]$ telle que :

$$\int_a^{x_o} f(x)\,dx = \int_{x_o}^{b} f(x)\,dx = \frac{1}{2}$$

mémoire centrale. Partie de l'unité centrale d'un calculateur constituée d'un grand nombre de registres (ou mots mémoires) destinés à stocker des informations; chaque mot mémoire est repéré par son adresse.

méromorphe. *adj.* **(fonction).** Se dit d'une fonction de variable complexe holomorphe sur un domaine D, à l'exception d'une partie de D sans points d'accumulation, formée par ses pôles.

mesure de densité. Une fonction réelle continue et positive p définit une mesure dite de densité p par rapport à la mesure de Lebesgue, par :

$$f \rightarrow \int f(x)\,p(x)\,dx.$$

Si m est une mesure positive sur X et p une fonction positive localement intégrable sur X, l'application $f \rightarrow \int f \cdot p \cdot dm$ est dite mesure de densité p pour m. C'est la généralisation de la dérivation.

mesure de Radon. Une mesure de Radon positive sur un ensemble X est une forme linéaire continue et positive sur l'espace vectoriel $\mathcal{K}$ (X) des fonctions continues sur X, nulles en dehors d'un compact de X.

méthode des moindres carrés. Méthode utilisée pour résoudre les problèmes de régression. On distingue la méthode des moindres carrés ordinaire et la méthode des moindres carrés généralisée.

métrique. *adj.* **(espace).** Se dit d'un ensemble muni d'une distance ou métrique.

métrisable. *adj.* Se dit d'un espace topologique tel qu'il existe une métrique qui engendre sa topologie. On dit encore espace *régulier.*

minimax ou **col.** *n.* U_{ij}, élément d'une matrice de jeu de rang $m \times n$ pour un jeu à deux personnes de somme nulle, est un minimax si :

$$U_{ij} \leqslant U_{ik}, \; k = 1, n$$
$$U_{ij} \geqslant U_{lj}, \; l = 1, m.$$

minorant. *n.* Dans un ensemble ordonné E, un minorant d'un sous-ensemble A est un élément m de E inférieur ou égal à tous les éléments de A.

mode. *n.* Le mode de la variable aléatoire X admettant la densité de probabilité $f(x)$ est la valeur x_o, si elle existe et si elle est unique, pour laquelle $f(x)$ est maximum.

moment. *n.* Le moment d'ordre n $(n \in N_+)$ de la variable aléatoire X est l'*espérance mathématique* de la variable aléatoire X^n.

moyenne. *adj.* **(valeur).** La valeur moyenne de la variable aléatoire X est l'*espérance mathématique* de X.

multinômiale. *adj.* **(loi).** La loi multinômiale de rang k est la loi de probabilité régissant le tirage répété n fois de la variable aléatoire discrète X pouvant assumer k valeurs distinctes. Si n_i est le nombre de résultats donnant la i-ème valeur x_i de X, alors :

$$\Pr(n_1, \dots n_k) = \frac{n!}{\prod_i n_i!} \prod_i p_i^{n_i}$$

où p_i est la probabilité d'obtenir x_i en un tirage et $\sum_i n_i = n$.

multiple. *n.* ou *adj.* Dans un anneau, b est un multiple de a s'il existe un élément k de l'anneau tel que $b = ka$. *Ex. :* dans $\mathbb{Z}$, 10 est un multiple de 5.

N-O

norme. *n.* Pour une variable aléatoire discrète, la norme N de la variable aléatoire est :

$$N = \sum_i F(x_i)$$

où F (x_i) est la loi de probabilité associée à X. Pour une variable aléatoire continue, définie sur l'intervalle $[a, b]$, la norme est :

$$N = \int_a^b f(x)\,dx$$

où $f(x)$ est la densité de probabilité associée à X.

noyau. *n.* Partie d'un espace vectoriel E dont tous les éléments ont une image, par une application linéaire f : E $\rightarrow$ F, égale au vecteur nul de F.

optimum. *n.* Soit f une application de E dans $\mathbb{R}$. Un élément x^* de E sera dit optimum de f s'il est un maximum ou un minimum de la fonction c'est-à-dire :

$$f(x) \leqslant f(x^*) \;\forall x \in E \text{ (maximum)}$$
$$f(x) \geqslant f(x^*) \;\forall x \in E \text{ (minimum).}$$

Si E est un espace topologique et si l'une ou l'autre de ces propriétés a lieu dans un voisinage U de x^*, on dira que l'on a affaire à un optimum local.

ordonnance. *n.* Ordre sur les paires d'éléments d'un ensemble.

ordre. *n.* Relation binaire sur un ensemble, réflexive, antisymétrique et transitive. Un ordre est total si deux éléments quelconques sont toujours comparables; sinon, il est partiel.

orientable. *adj.* Qualifie une surface admettant une décomposition en triangles (gauches) orientés, telle que l'ensemble des orientations n'admette pas de « contradiction ».

ouvert. *adj.* Qualifie un sous-ensemble d'un espace topologique tel que pour tout point de ce sous-ensemble il existe une boule ouverte ayant ce point pour centre, et contenue dans le sous-ensemble.

P

parabole. *n.* Une parabole est une conique dont l'équation peut se mettre sous la forme :

$$y^2 - 4\,cx = 0.$$

paraboloïde elliptique. Le paraboloïde elliptique est une quadrique qui peut se représenter par l'équation :

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 2\,z.$$

paraboloïde hyperbolique. Le paraboloïde hyperbolique est une quadrique qui peut se représenter par l'équation :

$$-\frac{x^2}{a^2} + \frac{y^2}{b^2} = 2\,z.$$

partie (d'un ensemble). Soit E un ensemble; tout ensemble A tel que A $\subset$ E est une partie de E.

partition (d'un ensemble). Recouvrement d'un ensemble par des parties disjointes.

période. *n.* On dit qu'une application f de $\mathbb{R}$ dans $\mathbb{R}$ admet la période T ou est périodique de période T ou encore est T-périodique si pour tout x réel on a :

$$f(x + T) = f(x).$$

périphérique. *n.* Organe extérieur à l'unité centrale et permettant à celle-ci d'échanger des informations avec le monde extérieur.

plan. *n.* L'espace étant rapporté à un système orthonormé, un plan est le lieu des points vérifiant une équation de la forme : $ax + by + cz + d = 0$.

pleine. *adj.* **(partie).** La partie pleine d'un ensemble est cet ensemble lui-même.

plus grand élément. Dans un ensemble ordonné E, un plus grand élément d'un sous-ensemble A est un élément de A supérieur ou égal à tous les éléments de A qui lui sont comparables.

plus petit élément. Dans un ensemble ordonné E, un plus petit élément d'un sous-ensemble A est un élément de A inférieur ou égal à tous les éléments de A qui lui sont comparables.

Poisson (loi de). Loi de probabilité F (n) associée à la variable aléatoire discrète X, définie sur N_+, telle que :

$$F(n) = \frac{e^{-m}\,m^n}{n!},$$

où $m = E(X) \in R_+^*$.

premier. *adj.* **(nombre).** Se dit d'un nombre entier relatif n'admettant pas d'autres diviseurs que lui-même et l'unité.

préordre (sur un ensemble). Relation binaire, réflexive et transitive.

primitive. *n.* La fonction G dérivable en tout point d'un intervalle I est une primitive de la fonction g sur I si sa dérivée en tout point de I est égale à g.

principe de récurrence ou **d'induction complète.** Ce principe dit que, si un ensemble d'entiers naturels S contient 0 et s'il contient $n+1$ à chaque fois qu'il contient n, alors S coïncide avec l'ensemble $\mathbb{N}$ de tous les entiers naturels.

probabilité intrinsèque. Grandeur objective caractérisant un événement aléatoire, telle que la fréquence d'apparition de cet événement dans une suite d'épreuves répétées en soit une approximation. Par extension mathématique, on peut appeler probabilité toute application P d'un espace probabilisable $(\mathcal{E}, E)$ dans $\mathbb{R}_+$ satisfaisant aux axiomes :

$$\forall a \in (\mathcal{E}, E) \quad 0 \leqslant P(a) \leqslant 1, \quad P(\mathcal{E}) = 1$$

et telle que pour tout couple d'événements exclusifs $a, b \in (\mathcal{E}, E)$:

$$P(a \cup b) = P(a) + P(b).$$

problème aux limites. Recherche des solutions d'une équation aux dérivées partielles, satisfaisant à une ou plusieurs conditions initiales (on dit encore *problème de Cauchy*).

programme. *n.* Ensemble des informations (instructions, valeurs numériques, réservations de zones mémoire pour stocker des données) nécessaires à la résolution d'un problème donné.

propre. *adj.* 1° *Sous-espace :* partie d'un espace vectoriel formée par tous les vecteurs propres ayant la même valeur propre associée. 2° *Valeur, vecteur :* un vecteur V colinéaire à son image par une application linéaire est dit vecteur propre de cette application ; le rapport de colinéarité est dit valeur propre associée.

propre. *adj.* **(partie).** Partie d'un ensemble qui n'est ni vide ni pleine.

puissance du continu. Cardinal d'un ensemble équipotent à l'ensemble $\mathbb{R}$ des nombres réels.

puissance d'un ensemble. Cardinal de cet ensemble.

Q-R

quadrique. *n.* Les quadriques sont les surfaces décrites par des coniques situées dans des plans parallèles lorsque leurs sommets décrivent d'autres coniques. C'est le cas de toute surface dont l'équation en coordonnées cartésiennes est du second degré en x, y et z.

racine carrée du nombre *a*. Racine positive du polynôme $X^2 - a$, où a est un nombre réel positif ; on la note $\sqrt{a}$.

raisonnement indirect. Exprimé symboliquement par la formule :

$$[(\neg p \to q) \wedge (\neg p \to \neg q)] \to p.$$

Il permet d'établir un énoncé p en montrant qu'à partir de la négation de p, on peut déduire des contradictions. Supposons, par exemple, qu'on veuille établir « qu'il existe un entier n ayant la propriété P. On partira de sa négation : « Aucun entier n n'a la propriété P. » Les contradictions qui s'en déduisent montrent « qu'il est faux qu'aucun entier n n'a la propriété P ». Jusqu'à ce point, les mathématiciens de toute obédience acceptent la conclusion. Mais les intuitionnistes se séparent des classiques en refusant de passer de cette conclusion à la suivante : « Il existe un entier n qui a la propriété P », tant qu'on n'a pas effectivement exhibé ou « construit » un tel entier. Le raisonnement indirect est donc un type de preuve non constructive, acceptée sans problèmes seulement dans la mathématique classique.

rang. *n.* Le rang d'une application linéaire, ou de sa matrice relativement à deux bases données, est la dimension du sous-espace image.

rationnel. *adj.* **(nombre).** Se dit d'un nombre qui peut être représenté par le quotient de deux nombres entiers.

rayon de convergence. Rayon du cercle du plan complexe tel qu'en tout point intérieur une série entière donnée soit convergente, et qu'en tout point extérieur elle soit divergente. On ne peut rien dire *a priori* du comportement de la série en un point de la circonférence de ce cercle.

réciprocité. *n.* Propriété de la transformation de Fourier, telle que la transformation de Fourier est la réciproque de sa conjuguée.

recouvrement. *n.* Un recouvrement ouvert (respectivement fermé) d'un sous-espace F d'un espace topologique E est une collection d'ouverts (respectivement de fermés) de E tels que leur réunion contienne F. En d'autres termes, les sous-ensembles $A_1, \ldots, A_n$ forment un recouvrement de l'ensemble E si leur réunion contient E.

redondance. *n.* Caractérise une information formellement superflue. Pratiquement, elle peut être utilisée pour des raisons de sécurité.

réduite. *adj.* **(variable aléatoire).** Soit X une variable aléatoire admettant une variance σ^2 (X). La variable aléatoire $Y = X/\sigma^2(X)$ est la variable aléatoire réduite associée à X.

registre. *n.* Association de plusieurs bistables permettant de stocker un mot ou un nombre binaire de plusieurs bits.

registre d'adresse mémoire. Registre dans lequel est chargée l'adresse d'un mot mémoire pour y être décodée avant une opération de lecture ou d'écriture.

registre de base. Registre dans lequel est chargée l'adresse de base d'une séquence d'instruction, en début de séquence, lorsque l'adressage dans le calculateur se fait par base et déplacement.

registre de données mémoire. Registre pour lequel transitent toutes les informations échangées entre la mémoire centrale et d'autres parties du calculateur.

registre d'index. Registre dont le contenu est ajouté à l'adresse spécifiée dans l'instruction lorsqu'il s'agit d'un adressage indexé.

registre d'instruction. Registre dans lequel est chargée l'instruction binaire pour y être décodée avant exécution.

règlement. *n.* Résultat chiffré d'une partie (théorie des jeux).

régression. *n.* Méthode statistique consistant à « expliquer » une variable (régression simple) ou plusieurs (régression multiple) en fonction d'autres variables.

régulière. *adj.* **(chaîne).** Chaîne d'événements aléatoires dont la matrice stochastique P de rang $m \times n$ est telle que :

$$\forall_i \ (1 \leqslant i \leqslant m) \quad \forall_j \ (1 \leqslant j \leqslant n)$$
$$\exists k \in \mathbb{N}^*_+ \quad (P^k)_{ij} \neq 0.$$

relation (entre deux ensembles). Relation définie par un triplet (E, F, G) où E est l'ensemble de départ, F l'ensemble d'arrivée et G le graphe de la relation, c'est-à-dire l'ensemble des couples (de E x F) pour lesquels la relation est vraie.

représentant (d'une classe d'équivalence). Un quelconque des éléments de cette classe. La connaissance d'un représentant d'une classe suffit à déterminer celle-ci.

résidu. *n.* Nom donné au coefficient du terme en $1/z$ dans le développement en série de Laurent d'une fonction de variable complexe au voisinage d'un point singulier.

résidu aléatoire. Variable aléatoire mesurant l'écart entre la variable à expliquer et une fonction préalablement choisie des variables explicatives.

résoluble par radicaux. Propriété d'une équation algébrique selon laquelle ses racines s'expriment à l'aide des quatre opérations et de l'extraction de racines m-ièmes.

réunion. *n.* Pour deux ensembles, c'est l'ensemble des éléments qui appartiennent à l'un ou à l'autre des ensembles. *Ex. :*

$$x \in A \cup B \Leftrightarrow [x \in A \textbf{ ou } x \in B].$$

Désigne également l' « opération » qui associe à deux ensembles leur réunion, notée ∪.

rupture de séquence. Passage d'une instruction à une autre instruction qui n'est pas forcément située à l'adresse suivante.

S

semi-norme. *n.* Une semi-norme sur un espace vectoriel E est une fonction $p : E \to \mathbb{R}^+$ ayant les propriétés suivantes :

1° $p(\vec{x}) \geqslant 0 ;\ p(\vec{0}) = 0$

2° $p(\lambda\vec{x}) = |\lambda|\, p(\vec{x}) \quad \lambda \in \mathbb{R}$

3° $p(\vec{x} + \vec{y}) \leqslant p(\vec{x}) + p(\vec{y})$.

Une norme est une semi-norme.

séparé. *adj.* Un espace topologique E est dit séparé si pour tout couple de points x et y de E il existe deux ouverts d'intersection vide, l'un contenant x et l'autre y.

séquentiel. *adj.* **(circuit).** Qualifie un circuit électronique logique, dont l'état dépend à la fois de l'état des entrées et des états précédents.

série. *n.* Une série numérique est définie par une suite (a_n) de nombres — avec $n \in \mathbb{N}$ et par la suite (S_n) $n \in \mathbb{N}$ telle que $S_p = a_0 + a_1 + \ldots a_p$ pour tout $p \in \mathbb{N}$ (suite des sommes partielles).

série entière. Une série entière est une série de fonctions dont le terme général est un monôme $a_n (z - z_0)^n$; on dit que z_0 est le centre de la série.

série de Fourier. Série trigonométrique associée à une fonction T-périodique f. Une des façons de l'exprimer est :

$$\sum_{n \in \mathbb{Z}} c_n e^{inx},$$

en posant :

$$c_n = \frac{1}{T} \int_0^T f(x) \cdot e^{-\frac{2\pi}{T} inx}\, dx.$$

signe, sens, dénotation. Selon F. de Saussure, un *signe* est un phénomène double où un signifiant (vocal, écrit, gestuel) est relié à un signifié qui est la contrepartie conceptuelle du signifiant. Le signifié d'un signe est solidaire d'une langue donnée, naturelle ou artificielle.
Le *sens,* lui, est un contenu de pensée (voir les quelques indications sur Frege dans le texte *Logique*) essentiellement susceptible d'être traduit par d'autres signes dans une autre langue. Mais le sens d'une expression (suite de signes) ne se livre que de façon différentielle, c'est-à-dire par rapport à et en fonction du contexte de l'expression. L'analyse du sens d'une expression n'implique aucun renvoi du langage à quelque chose d'extérieur à lui et se maintient dans l'enceinte du discours.
Mais le discours se rapporte aussi aux choses, réalités extralinguistiques, et ce rapport, désigné comme *dénotation* ou référence, institue la question de la valeur de la vérité d'une expression.

similarité (indice de). Indice de mesure de proximité, utilisé dans les méthodes statistiques de classification.

similitude. *n.* Correspondance biunivoque entre deux figures géométriques, telle que deux segments homologues soient toujours proportionnels avec le même rapport k, appelé rapport de similitude.

simple. *adj.* **(extension).** Se dit d'une extension L d'un corps K telle qu'il existe $l \in L$ tel que L soit le corps K (l) engendré par K et l.

sous-anneau. *n.* Un sous-anneau B d'un anneau A est une partie B de A qui est elle-même un anneau pour les lois de A induites sur B.

sous-corps. *n.* Un sous-corps L d'un corps K est une partie L de K qui est elle-même un corps pour les lois induites sur L de celles de K.

sous-ensemble. *n.* Un sous-ensemble A de l'ensemble E est un ensemble tel que $A \subset E$ (voir aussi *partie d'un ensemble*).

sous-groupe. *n.* Si G est un groupe, une partie H de G est un sous-groupe de G si H est un groupe pour la loi induite de celle de G. H sera un sous-groupe de G si et seulement si la propriété suivante est vérifiée : si $x \in H$ et $y \in H$, alors $xy^{-1} \in H$.

sous-groupe distingué. Un sous-groupe H d'un groupe G est un sous-groupe distingué (ou invariant) de G si, pour tout élément x de G, on a :

$$xHx^{-1} = H.$$

sous-programme. *n.* Partie de programme, pouvant être exécutée plusieurs fois, à la suite d'appels pouvant figurer en n'importe quel point du programme. Le sous-programme n'est stocké qu'une fois, même s'il doit être exécuté plusieurs fois.

spectre. *n.* Ensemble de toutes les valeurs propres d'une application linéaire.

sphère. *n.* Une sphère est, dans l'espace, le lieu des points situés à une distance r constante d'un point fixe C (α, β, γ). Son équation est :

$$(x - \alpha)^2 + (y - \beta)^2 + (z - \gamma)^2 = r^2.$$

sphérique. *adj.* 1° *Angle :* angle plan que font deux tangentes à deux côtés d'un triangle sphérique en leur point d'intersection. 2° On désigne par *excès sphérique* la quantité dont la somme des angles d'un triangle sphérique dépasse celle des angles d'un triangle plan. 3° Un *triangle sphérique* est la partie d'une surface sphérique délimitée par trois arcs de grand cercle (chacun des arcs étant plus petit qu'un grand demi-cercle).

stochastique. *adj.* **(processus, modèle).** Le calcul classique des probabilités s'applique à des épreuves où chaque résultat possible est un nombre. Cependant, bien des situations réelles relèvent de modèles aléatoires de nature plus complexe, par exemple l'évolution d'une population donnée en fonction du temps.
— Une probabilité p est une application d'une classe $\mathcal{B}$ de parties d'un ensemble Ω ($\mathcal{B}$ ayant les propriétés d'une tribu), sur le segment fermé [0, 1] (la probabilité d'un événement *certain* est évidemment égale à 1). On appelle espace de probabilité le triplet (Ω, $\mathcal{B}$, p).
— Une variable aléatoire à valeurs dans Ω_1 est une application (mesurable) d'un espace de probabilité (Ω, $\mathcal{B}$, p) dans Ω_1 muni d'une tribu $\mathcal{B}_1$.
— Un *processus stochastique* est une famille de variables aléatoires, c'est-à-dire qu'à tout temps t est associée une variable aléatoire prenant ses valeurs dans un ensemble numérique. Un processus stochastique est ainsi une fonction aléatoire dont l'argument est le temps, dont le déroulement est à la fois irréversible et inéluctable.

stratégie. *n.* Ensemble de règles déterminant la conduite d'un joueur, compte tenu de toutes les réactions possibles de l'adversaire (ou des adversaires), jusqu'au règlement final de la partie.

stratégie dominante. Stratégie A_k telle que, quelle que soit la stratégie B_i choisie par l'adversaire, elle ne conduise pas à un règlement inférieur à une autre stratégie A_j. La matrice de jeu U de rang $m \times n$ est donc telle que :

$$\forall i \ (1 \leqslant i \leqslant n)$$
$$\forall j \ (1 \leqslant j \leqslant m), \ U_{ki} \geqslant U_{ji}.$$

stratégie mixte. Manière de conduire un jeu telle que, au cours d'une série répétée de parties, la stratégie appliquée varie d'une partie à l'autre de façon aléatoire, la stratégie A_i ayant une probabilité définie P_i d'être appliquée.

suite. *n.* En mathématiques, une suite est une application de l'ensemble $\mathbb{N}$ des nombres entiers naturels dans un ensemble quelconque E. Cette application associe à chaque entier n de $\mathbb{N}$ un élément e (unique) de E d'où la notation : $e_0, e_1, e_2, \ldots, e_n, \ldots$

suite de choix. Ce concept est dû à Brouwer. On le définira sur l'exemple simple des nombres naturels. Une suite α est obtenue par des choix successifs, tels qu'on puisse imposer au cours du processus des restrictions sur les choix ultérieurs. C'est donc une suite de couples (a_i, R_i), où R_i représente les conditions sur les suites d'entiers ; ces conditions se restreignent progressivement, c'est-à-dire que $R_i\alpha$ implique $R_{i+n}\alpha$. Les suites de choix ne sont pas des objets achevés ; à chaque instant, on n'en connaît qu'un segment initial.

supplémentaires. *n.* Sous-espaces vectoriels d'un espace vectoriel E n'ayant que le vecteur nul en commun et engendrant E par addition.

sur-corps. *n.* Un corps K est dit sur-corps d'un corps L si L est un sous-corps de K.

symbolique. *adj.* **(calcul).** Méthode de résolution des équations de convolution partant du fait que la transformée de Laplace d'un produit de convolution est égale au produit simple des transformées. On change alors de structure algébrique de référence.

système d'exploitation. Ensemble de programmes destinés à l'exploitation la plus facile possible du calculateur par le plus grand nombre possible d'utilisateurs, ceux-ci n'étant pas forcément des spécialistes du matériel utilisé.

système fondamental de voisinages. Soit $(V_i)_{i \in I}$ une famille de voisinages d'un élément a d'un espace topologique E. On dit qu'il s'agit d'un système fondamental de voisinages si tout voisinage de a contient l'un des V_i.

système de référence cartésien (dans le plan et l'espace). Un système cartésien est la donnée de deux (dans le plan) ou de trois (dans l'espace) droites orientées se coupant en un point O (origine), sur chacune desquelles est fixée un vecteur unitaire.

système orthonormé. Un tel système est un système cartésien dans lequel les vecteurs unitaires sont de longueurs égales et les axes orthogonaux.

T-U-V-W-Z

table de vérité. Représentation des fonctions logiques par une table où figurent toutes les intersections de base et la valeur correspondante de la fonction.

tableau de contingence. Tableau construit de telle sorte qu'à l'intersection de la ligne i et de la colonne j on trouve le nombre d'individus possédant à la fois les caractères i et j. Un tel tableau se prête généralement bien à une analyse des correspondances.

temps réel. Mode d'utilisation d'un calculateur, tel que celui-ci travaille au rythme d'un phénomène extérieur pour acquérir des données concernant ce phénomène et les traiter au fur et à mesure qu'elles arrivent.

tenseur (en dualité). Un tenseur covariant d'ordre p sur un espace vectoriel E est une forme p-linéaire sur l'espace E. Un tenseur contravariant d'ordre q sur E est une forme q-linéaire sur l'espace dual E^*.

tensoriel. *adj.* **(produit).** Forme multilinéaire de plusieurs tenseurs, qui prend pour valeur le produit des valeurs prises par ces tenseurs. Le produit tensoriel peut être contracté en opérant, *par ex.*, une contraction d'indices sur les composantes du produit.

terme. *n.* En logique aristotélicienne, on désigne par là les éléments autres que la copule qui composent la phrase type construite sur le moule : « Socrate est mortel » ; dans cet exemple *Socrate* et *mortel* sont des termes ; les termes sont donc les sujets ou prédicats des propositions.
Au sens technique de la logique mathématique contemporaine, « terme » se définit par récurrence selon les règles (1) et (2) :
(1) Une variable ou une constante est un terme ;
(2) Pour un symbole de fonction f et des termes $t_1, \ldots t_m$, $f\,(t_1, \ldots t_m)$ est un terme (voir texte *Logique*,).

théorème de Bolzano-Weierstrass. Il s'énonce comme suit : « Tout ensemble de nombres réels infini borné admet au moins un point d'accumulation », ou encore : « De toute suite de réels bornée on peut extraire au moins une suite partielle convergente. » La démonstration de ce théorème fondamental de l'analyse repose sur la loi du tiers exclu, et c'est pourquoi il tombe sous la critique des intuitionnistes.

théorème de Zorn. Théorème dont l'énoncé est le suivant : tout ensemble ordonné inductif admet un élément maximal.

théorie de la connaissance. On appelle ainsi toute théorie philosophique qui s'interroge sur la nature et la portée de la connaissance humaine, sur son degré d'objectivité et la mesure dans laquelle nous pouvons ajouter foi à ce qu'elle prétend nous apprendre sur la réalité. En demandant ce qu'est la connaissance, par opposition à la croyance et à l'opinion, et en mettant en doute l'apport des sens, Platon a élaboré la première théorie de la connaissance.

théorie de la réminiscence. C'est la plus vieille théorie qui correspond à celle, plus moderne, de l'*a priori*. Partant de l'impossibilité de donner une explication expérimentale satisfaisante et cohérente aux vérités mathématiques et, de manière générale, à toutes les propriétés intelligibles, Platon établit cette doctrine pour montrer que la mathématique n'est que la reproduction ou la répétition assourdie de souvenirs, la réeffectuation d'un savoir mathématique séparé du monde sensible, inaccessible aux contingences du devenir et du temps.

théorie des « visions du monde ». Le terme allemand, dû à Dilthey, est « Weltansehbauungen ». Dilthey pense que l'homme a une indéracinable tendance à forger des interprétations globales de la réalité, des « visions du monde », c'est-à-dire des systèmes plus ou moins cohérents d'idées et de représentations, de principes éthiques et d'expressions religieuses et artistiques. Ces systèmes, subjectifs et relatifs, constituent néanmoins un aspect authentique de la vie.

topologie. *n.* Un ensemble E est muni d'une topologie si on a défini sur E une famille de parties, appelées parties ouvertes, vérifiant les propriétés suivantes :
— E et $\varnothing$ sont des ouverts ;
— toute réunion (finie ou infinie) d'ouverts est un ouvert ;
— toute intersection finie d'ouverts est un ouvert.

topologie affaiblie du primal. On appelle topologie affaiblie de E la topologie d'espace vectoriel définie par la famille de semi-normes :

$$P_{A'}\,(\vec{e}) = \sup |< e', e >| \ \overleftarrow{e'} \in A'$$

A′ partie finie de E′.

topologie faible du dual. On appelle topologie faible de E′ la topologie d'espace vectoriel semi-normé, définie par la famille de semi-normes :

$$P_A\,(\overleftarrow{e'}) = \sup |< e, e' >| \ \vec{e} \in A$$

A partie finie de E.

topologie forte du dual. Soit E un espace vectoriel normé, E′ son dual topologique. On appelle topologie forte du dual la topologie définie par la norme :

$$\|\overleftarrow{e'}\| = \sup |< \overleftarrow{e'}, \vec{e} >| \ \ \|\vec{e}\| \leqslant 1.$$

transcendant. *adj.* **(nombre).** Nombre qui n'est racine d'aucun polynôme à coefficients entiers ; e, π sont des nombres transcendants.

transcendante. *adj.* **(extension).** Extension simple K (l) d'un corps K telle qu'aucun polynôme non nul à coefficients dans K n'admette l pour racine.

transfini. *adj.* **(nombre).** Cardinal d'un ensemble contenant un ensemble équipotent à l'ensemble $\mathbb{N}$ des entiers naturels.

tribu. *n.* Famille de parties d'un ensemble, formant un clan, telle qu'une réunion dénombrable de parties soit encore dans la famille.

ultramétrique. *n.* Distance (ou métrique) particulière utilisée en classification.

union (de deux ensembles). Voir *réunion*.

unité arithmétique et logique. Ensemble de circuits réalisant les opérations de types arithmétique et logique.

unité centrale. Constituée de l'unité arithmétique et logique, de l'unité de contrôle et de la mémoire centrale, elle constitue un ensemble indépendant pouvant exécuter des programmes. Elle communique avec le monde extérieur par les périphériques.

unité de contrôle. Ensemble de circuits assurant le bon déroulement des opérations. L'unité de contrôle reçoit les instructions une à une, les décode et les fait exécuter.

utilité. *n.* Règlement d'un jeu tel qu'une transformation linéaire appliquée à l'ensemble des règlements n'affecte pas la conduite des joueurs.

valeur absolue (d'un nombre réel). x étant un nombre réel, on appelle valeur absolue de x et on note $|x|$ le nombre réel positif égal à x si x est positif et à $-x$ si x est négatif.

variance. *n.* Soit X une variable aléatoire. On appelle variance σ^2 associée à X l'espérance mathématique du carré de la variable centrée correspondante :

$$\sigma^2 = E(X - E(X))^2.$$

vide. *adj.* **(ensemble** ou **partie).** C'est l'ensemble qui n'a aucun élément, noté $\varnothing$. *Ex. :* $\forall x, x \notin \varnothing$. La mathématique n'a pas horreur du vide : on rencontre souvent l'ensemble vide, mais il n'y en a qu'un.

voisinage. *n.* Toute partie d'un espace topologique E contenant au moins un ouvert contenant lui-même un point $a \in E$ est dite voisinage de a.

NOTE RELATIVE AU GÉNÉRIQUE DU VOLUME XIII

Ont collaboré à ce volume :

M. BELLEC, pour les ensembles, les nombres.
M. BELLEC et C. PARDOUX, pour l'analyse de données.
B. GOLDFARB et J. THÉPOT, pour l'introduction aux mathématiques.
B. GOLDFARB, pour l'algèbre linéaire, les équations algébriques, l'analyse, le calcul tensoriel.
R. HARA, pour le langage ensembliste.
R. LESTIENNE, pour statistiques et probabilités.
C. PARDOUX, pour la géométrie, la trigonométrie, la topologie, les courbes et surfaces.
Y. RIO, pour l'introduction à l'informatique.
H. SINACEUR, pour mathématiques et philosophie.
A. et H. SINACEUR, pour logique.
A. SINACEUR, pour mathématiques et société, mathématiques et pédagogie, histoire des mathématiques.
C. SOFFER-SACOTTE, pour les structures algébriques, la géométrie analytique.
J. THÉPOT, pour l'analyse combinatoire, les graphes, l'analyse fonctionnelle, le calcul numérique, les mathématiques financières et les mathématiques économiques.

ERRATA

Page 39, dans le paragraphe « la divisibilité », 6e et 7e lignes, lire : — il existe dans $\mathbb{Z}$ un élément neutre pour la multiplication, le nombre 1, tel que $a1 = 1a = a$.

Page 181, légende en bas, lire :

▼ Figure 10 : distribution de probabilité d'observer en une seconde, k désintégrations d'une substance radio-active subissant en moyenne 10 désintégrations par seconde.

SYMBOLES MATHÉMATIQUES USUELS

$\Rightarrow$	*implique*
$\Leftrightarrow$	*équivalent à*
$\forall$	*quel que soit... (ou : pour tout...)*
$\exists$	*il existe au moins un...*
$R \vee S$	*disjonction* (R *ou* S)
$R \wedge S$	*conjonction* (R *et* S)
$\neg R$	*négation (non* R*)*
$[a, b]$	*intervalle fermé d'origine a d'extrémité b*
$]a, b]$	*intervalle semi-ouvert à gauche*
$[a, b[$	*intervalle semi-ouvert à droite*
$]a, b[$	*intervalle ouvert*

ENSEMBLES DE NOMBRES

$\mathbb{N}$	*ensemble des entiers naturels*
$\mathbb{N}^*$	*ensemble des entiers naturels,* 0 *exclu*
$\mathbb{D}$	*ensemble des nombres décimaux*
$\mathbb{Q}$	*ensemble des nombres rationnels*
$\mathbb{Q}^*$	*ensemble des nombres rationnels,* 0 *exclu*
$\mathbb{R}$	*ensemble des nombres réels*
$\mathbb{R}^*$	*ensemble des nombres réels,* 0 *exclu*
$\mathbb{C}$	*ensemble des nombres complexes*

SYMBOLES DE LA THÉORIE DES ENSEMBLES

$\varnothing$	*ensemble vide*
$\{a, b\}$	*ensemble ayant pour éléments a et b*
Card E	*cardinal de l'ensemble* E
$\in$	*appartient à (ou : est élément de)*
$\notin$	*n'appartient pas à*
$\subset$	*est inclus dans*
$\not\subset$	*n'est pas inclus dans*
$\cup$	*réunion*
$\cap$	*intersection*
$\bigcup\limits_{i \in I} E_i$	*réunion d'un ensemble d'ensembles (réunion de la famille* E_i *pour i appartenant à* I*)*
$\bigcap\limits_{i \in I} E_i$	*intersection d'un ensemble d'ensembles*
$E \times F$	*produit cartésien de deux ensembles* E *et* F (E *croix* F)
(x, y)	*couple*
(x, y, z)	*triplet*
$(x_1, x_2, \ldots x_n)$	*n-uplet*
$\prod\limits_{i \in I} E_i$	*produit cartésien de la famille* E_i *pour i appartenant à* I
E^n	*produit cartésien de n ensembles égaux à* E

$E \triangle E$	*différence symétrique de deux ensembles*
$\complement_E F$	*complémentaire de* F *dans* E
$\mathcal{P}(E)$	*ensemble des parties d'un ensemble* E

RELATIONS

$x \equiv y \ (mod.\ \mathcal{R})$	*x congru à y modulo* $\mathcal{R}$
$E/\mathcal{R}$	*ensemble quotient de l'ensemble* E *par la relation d'équivalence* $\mathcal{R}$
$\preccurlyeq \quad \leqslant$	*relation d'ordre*
$\prec \quad <$	*relation d'ordre strict*
$x \mathcal{R} y$	*x est lié à y par la relation* $\mathcal{R}$
$f : x \mapsto f(x)$	*x a pour image f (x) par l'application f*
$f : E \to F$ ou $E \xrightarrow{f} F$	*application f d'un ensemble* E *dans un ensemble* F
Id E	*application identique de* E
f^{-1}	*application réciproque de l'application f*
$f \circ g$	*composé des applications f et g (f rond g)*

AUTRES SYMBOLES

$\sum\limits_{i=1}^{n} x_i$	*somme des n éléments* $x_1, x_2, \ldots x_n$
$\prod\limits_{i=1}^{n} x_i$	*produit des n éléments* $x_1, x_2, \ldots, x_n$
$\frac{df}{dx}$ *ou* $f'(x)$	*dérivée de f*
df	*différentielle de f*
$\frac{\partial f}{\partial x_i}$	*i-ième dérivée partielle de f*
dfx_0 *ou* dx_0f	*différentielle de f au point* x_0
$\int f(x)\,dx$	*intégrale de f.* $\iint$ *intégrale double.*
$\lvert x \rvert$	*valeur absolue de x ou module de x*
$\lVert x \rVert$	*norme de x*
$n!$	*factorielle n*
$\alpha, \beta, \lambda, \mu$	*scalaires*
Ω	*ensemble d'opérateurs*
grad. f	*gradient de la fonction f*
div. V	*divergence du champ de vecteurs* V
rot. V	*rotationnel du champ de vecteurs* V

INDEX DES NOMS CITÉS

Les références sont données par l'indication du numéro de la page où se trouve le terme, suivi, le cas échéant, des lettres *a* ou *b* se rapportant respectivement à la colonne de gauche ou à la colonne de droite de chaque page.

Les références sont données, pour les illustrations, par le numéro de la page en caractères gras et, pour le texte, par le numéro de la page en caractères maigres pour une simple citation, et en italique pour un développement plus complet.

Pour les différentes variantes se rattachant à un même terme, un astérisque indiquera le renvoi à ce dernier. *Ex. :*

compteur ordinal
* C.O.
C.O. (compteur ordinal) 246 a

A

B

C

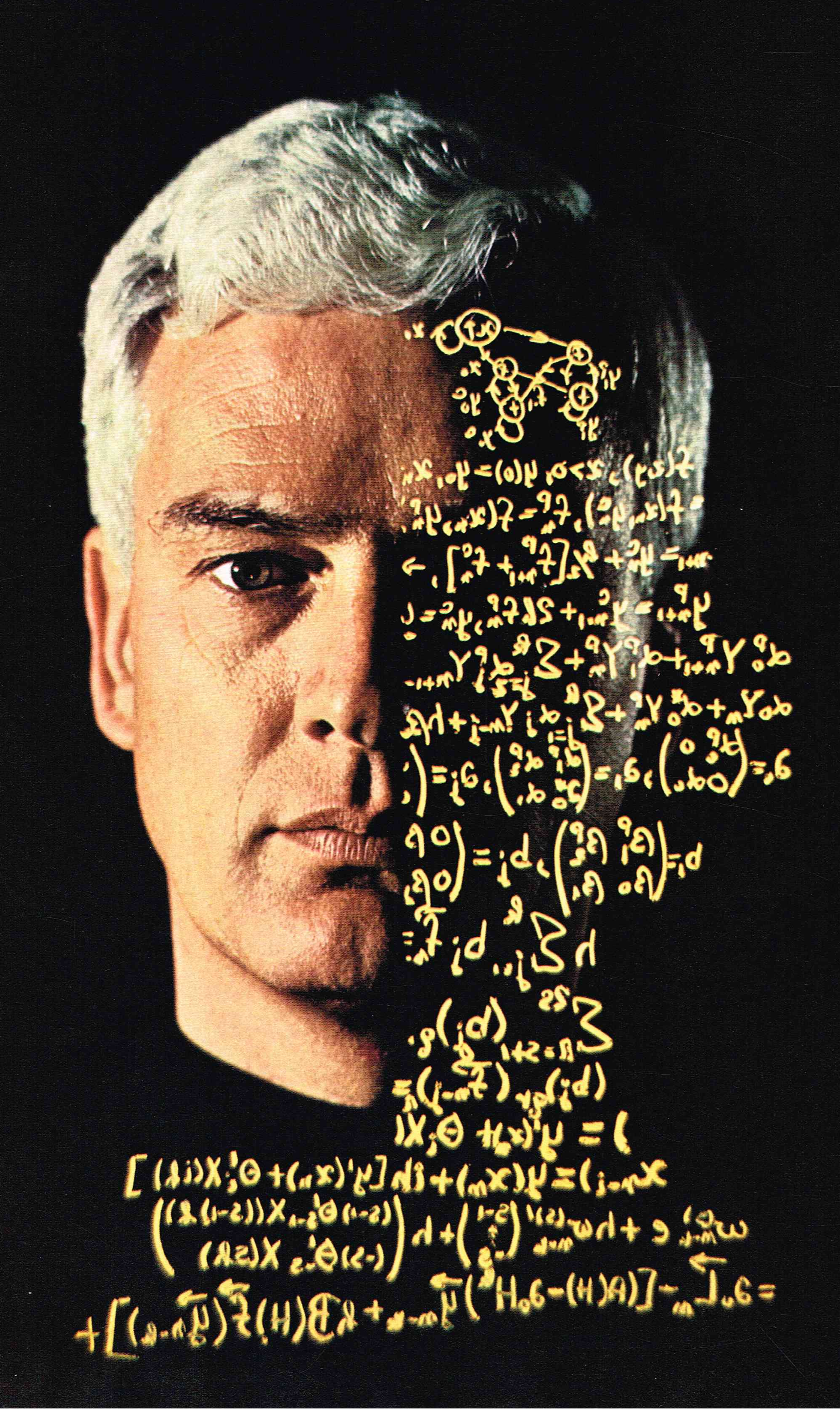
D

E

GRANDE ENCYCLOPÉDIE ALPHA DES SCIENCES ET DES TECHNIQUES

MATHÉMATIQUES

GRANGE BATELIÈRE, ÉDITIONS ATLAS, PARIS - ÉDITIONS KISTER, GENÈVE
ÉDITIONS ÉRASME, BRUXELLES-ANVERS



Publiée sous le haut patronage de :
Messieurs les professeurs :
Jean DORST, membre de l'Institut,
Charles FEHRENBACH, membre de l'Institut,
Roger HEIM, membre de l'Institut
Monsieur l'amiral André JUBELIN,
Messieurs les professeurs :
Pierre LÉPINE, membre de l'Institut,
Louis LEPRINCE-RINGUET, de l'Académie française,
Jean-François LEROY, professeur au Muséum national d'histoire naturelle,
Henri NORMANT, membre de l'Institut,
Monsieur Jacques PICCARD, docteur ès sciences h.c.

Ont collaboré à ce volume :

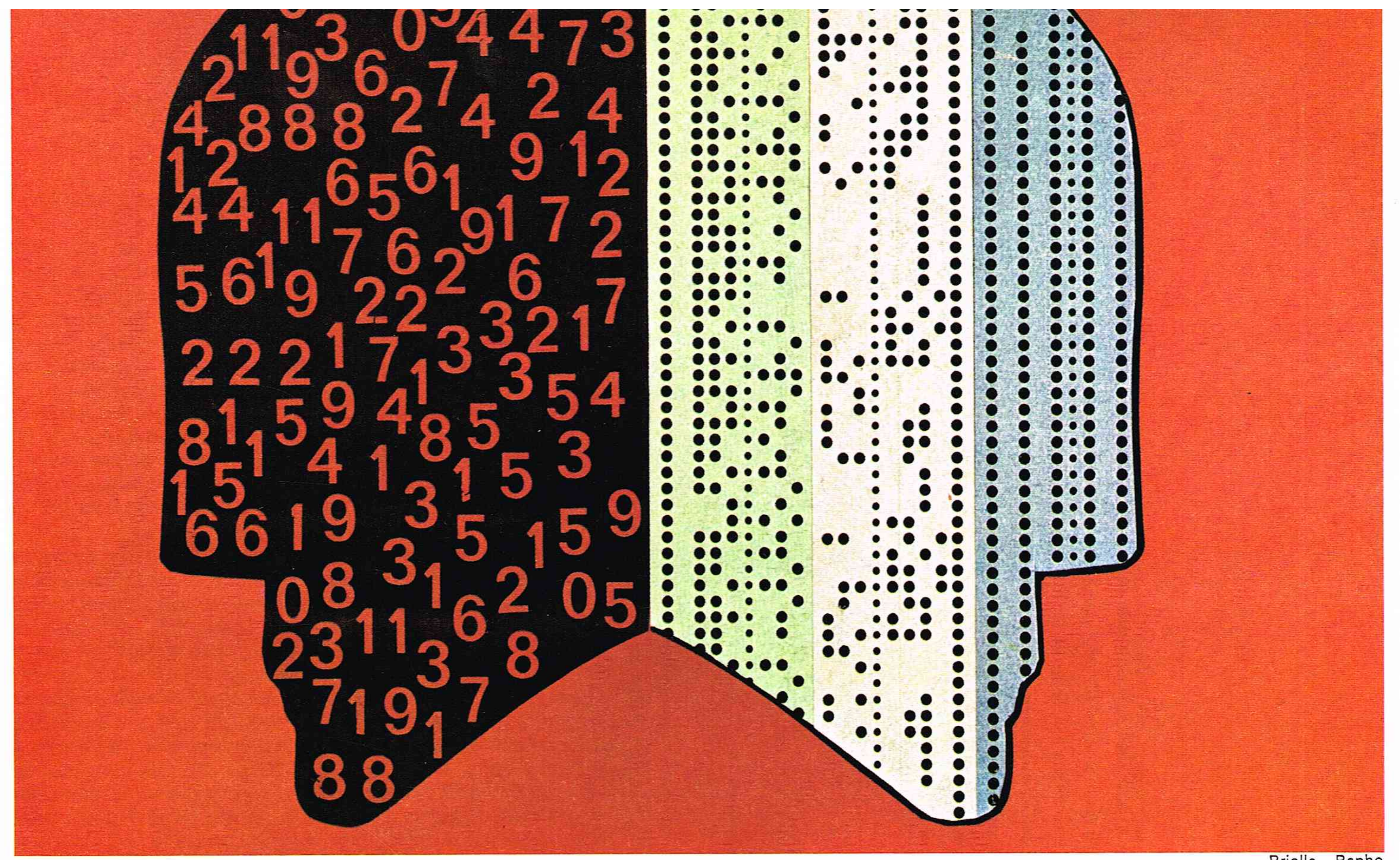
M. BELLEC, pour les ensembles, les nombres.
M. BELLEC et C. PARDOUX, pour l'analyse de données.
B. GOLDFARB et J. THÉPOT, pour l'introduction aux mathématiques.
B. GOLDFARB, pour l'algèbre linéaire, les équations algébriques, l'analyse, le calcul tensoriel.
R. HARA, pour le langage ensembliste.
R. LESTIENNE, pour statistiques et probabilités.
C. PARDOUX, pour la géométrie, la trigonométrie, la topologie, les courbes et surfaces.
Y. RIO, pour l'introduction à l'informatique.
H. SINACEUR, pour mathématiques et philosophie, logique, mathématiques et société, mathématiques et pédagogie, histoire des mathématiques.
C. SOFFER-SACOTTE, pour les structures algébriques, la géométrie analytique.
J. THÉPOT, pour l'analyse combinatoire, les graphes, l'analyse fonctionnelle, le calcul numérique, les mathématiques financières et les mathématiques économiques.

Les schémas portant la référence Richard Colin ont été réalisés d'après des croquis fournis par les auteurs.

Réalisation	IDÉES ET ÉDITIONS 16, avenue de Friedland, 75008 Paris
Comité de direction	Simone DEVAUX, Uberto TOSCO.
Rédaction	Patrick PHLIPONEAU, Françoise MENU, Marie-Noëlle PAILLETTE, Vanina DORÉ, Monique LIONS-GENTIL.
Recherche de l'illustration	Mathilde RIEUSSEC.
Mise en pages	Tito TOPIN et Serge BROCHE.
Illustrations techniques	Richard COLIN.
Coordinateur des dessins	Mario LOGLI.
Fabrication	Sylvia COLIN, Jocelyne TÉPÉNIER.
Directeur de la publication	G. BORDES.

Dans ce volume :

LES MATHÉMATIQUES



Briolle - Rapho

▲ ***La carte perforée, élément indispensable, de nos jours, de l'utilisation du calculateur électronique, a été inventée, sous une forme moins élaborée, il y a 250 ans ; on s'en servit pour la première fois sur des métiers à tisser.***

LES MATHÉMATIQUES

Au cours de ces cinquante dernières années, la part consacrée aux mathématiques dans les enseignements des lycées et des collèges n'a cessé d'augmenter. Impossible donc aujourd'hui d'espérer revêtir la robe de l'avocat, la blouse du chirurgien ou le tablier du charcutier sans avoir, des années durant, endossé la toge prétexte du mathématicien. Qu'il semble lointain le temps des humanités grecques et latines qui permettait à tel fils d'instituteur de province, lauréat du concours général de thème grec et grand amateur de poésie, d'accéder aux plus hautes fonctions politiques !

Cette évolution est due, à l'évidence, au développement des sciences et des techniques et à leur influence déterminante sur la croissance économique de nos pays. L'omniprésente informatique, par exemple, est grande consommatrice de mathématiques, et les progrès considérables réalisés en ce domaine rendent nécessaire la formation d'un grand nombre de personnes possédant un bon niveau en mathématiques.

Il faut cependant noter que cette évolution a également coïncidé (ce n'est, bien sûr, pas un hasard) avec un développement sans précédent de la science mathématique elle-même. Les programmes de l'enseignement ont été largement modifiés, non sans susciter des réactions parfois vives à l'encontre de ces mathématiques dites modernes. Il est bon de préciser cependant que ces mathématiques ne sont pas aussi modernes qu'on veut bien le dire ! En tant que telles, elles ont plus d'un siècle (ce qui est un âge fort raisonnable), puisqu'elles sont le point d'aboutissement d'un mouvement qui débute avec les travaux de Galois, de Cauchy et de Cantor, et date du siècle dernier. De plus, loin de sacrifier aux attraits de la modernité, ce mouvement est inspiré par les principes du classicisme le plus pur, car il est essentiellement l'expression d'un retour aux sources auxquelles la science mathématique va puiser à la fois son autonomie par rapport aux autres sciences et son unité. Autant dire que le qualificatif « moderne » est usurpé à plus d'un titre (comme toujours d'ailleurs).

La méthode

Où donc les mathématiques ont-elles trouvé leur unité, sinon au cœur même de la démarche intellectuelle que l'on retrouve aussi bien en analyse, en algèbre, en arithmétique qu'en géométrie ? Cette démarche est connue sous le nom de « méthode axiomatique ».

La ***méthode axiomatique*** *est généralement définie de la manière suivante : « De certaines propositions, appelées axiomes et supposées vraies, on déduit d'autres propositions par un enchaînement rigoureux, c'est-à-dire gouverné selon les règles de la logique formelle, elle-même soigneusement formalisée et axiomatisée. » Ainsi donc, si l'on admet cette définition, la méthode axiomatique serait purement déductive. Il s'agit là d'une restriction très couramment admise, qui peut conduire à de graves erreurs pédagogiques. A cet égard, Nicolas Bourbaki, mathématicien polycéphale, dit fort bien : « Ce que se propose pour but essentiel l'axiomatique, c'est précisément ce que le formalisme logique, à lui seul, est incapable de fournir, l'intelligibilité profonde des mathématiques [...]. La méthode axiomatique trouve son point d'appui dans la conviction que, si les mathématiques ne sont pas un enchaînement de syllogismes se déroulant au hasard, elles ne sont pas davantage une collection d'artifices plus ou moins astucieux faits de rapprochements fortuits où triomphe la pure habileté technique. »*

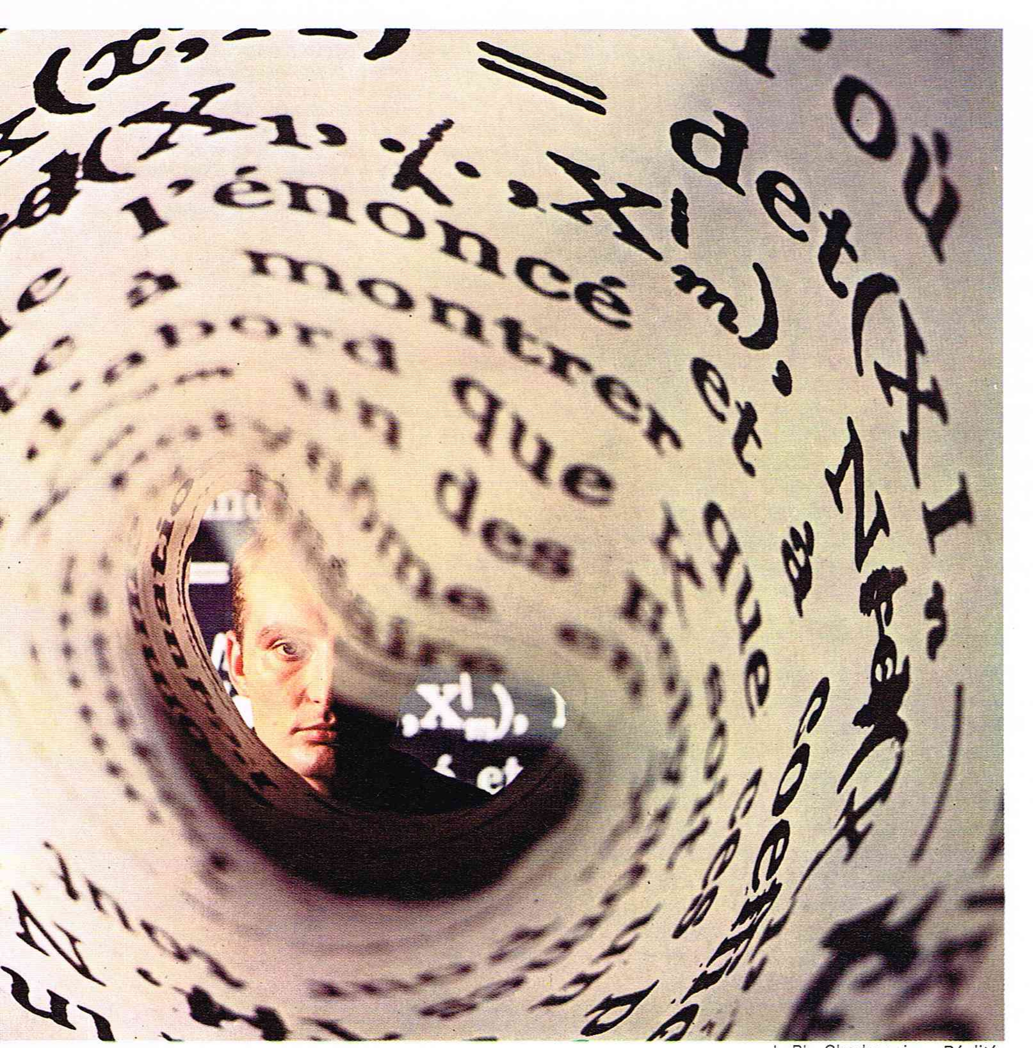
En réalité, la méthode axiomatique se décompose en deux phases distinctes, bien qu'étroitement complémentaires : une phase inductive et une phase déductive.

Au cours de la phase inductive, *le mathématicien donne libre cours à son imagination et à son intuition afin de déterminer l'énoncé qu'il désire établir. Pour cela, il se fiera à son propre jugement ainsi qu'à son expérience, c'est-à-dire à ce qu'il connaît du développement historique de son champ d'étude. Souvent, il s'agira de généraliser un résultat déjà connu en escomptant une certaine régularité des concepts étudiés. Allant du particulier au général, le mathématicien tente de résoudre ainsi la question : Que démontrer ? Suivra alors la* phase déductive, *au cours de laquelle il enchaînera les propositions d'une manière rigoureuse à partir des prémisses soigneusement formulées : « Le terrain que l'intuition a conquis ainsi d'un seul bond, il reste ensuite à l'organiser, à bâtir maillon par maillon la chaîne de propositions qui aboutira au résultat cherché. » (J. Dieudonné) Allant maintenant du général au particulier, le mathématicien tente de répondre à la question : Comment démontrer ? En réalité, ces deux phases sont étroitement liées. Elles se succèdent à la faveur d'un mouvement dialectique au cours duquel les concepts mathématiques sont définis et s'enrichissent dans un processus de synthèse permanent.*

Au bout du compte, seul subsiste le résultat de la phase déductive, c'est-à-dire une écriture de propositions agencées de manière cohérente. Mais ce résultat doit sa portée mathématique au fait qu'il a émergé d'un processus intellectuel évoluant de manière dialectique. Ainsi, contrairement à l'image qu'elle donne d'elle-même, la démarche mathématique ne se déroule pas de manière aveugle et mécanique dans un univers sans âme. A en croire le sens commun, le mathématicien serait semblable à une sorte d'explorateur des forêts tropicales qui se fraye un chemin rectiligne à travers une nature putride et inhospitalière, regardant droit devant lui sans jamais se retourner, sans risquer un coup d'œil sur ce qui l'entoure. Une telle image est bien imparfaite et ne rend guère compte de l'originalité de l'activité du mathématicien. La comparaison serait plus acceptable dans un univers surréaliste. L'explorateur est en communion intense avec la forêt qu'il pénètre ; chacun de ses pas modifie le relief du sol, la forme des arbres et la couleur de leurs feuilles. A chaque instant, il constate que le chemin parcouru s'est profondément transformé ; il ne vient pas d'où il est parti. Subitement pourvu du don d'ubiquité, voici qu'il réalise qu'il progresse dans plusieurs directions à la fois, bernant ainsi la rose des vents métamorphosée en girouette.

Restreindre la démarche mathématique à sa phase inductive, c'est, sous couleur de spontanéisme soi-disant génial, faire preuve de laxisme et de facilité en se contentant de vagues heuristiques ; de même, la restreindre à sa phase déductive, c'est la condamner à la stérilité et à l'ésotérisme dans une contemplation aveugle de truismes plus ou moins tautologiques.

▼ Il n'est pas possible de lire des mathématiques comme on lit un roman...



J.-Ph. Charbonnier - Réalités

Des ensembles

Appliquant ainsi au fil des siècles la méthode axiomatique — même sans trop se le dire — les mathématiciens ont progressivement conduit leur science à un point d'évolution où ils ont vivement ressenti la nécessité de dégager ce qui constituait la trame des diverses branches des mathématiques qu'ils distinguaient jusque-là.

La théorie des ensembles est ainsi apparue au cours du XIXe siècle. Sorte de théorie du langage mathématique, elle fournit une représentation claire et maniable des règles de logique élémentaire. Les notions d'ensemble, de sous-ensemble, d'application sont définies de manière formelle, ainsi que les règles opérationnelles que l'on connaît (réunion, intersection...) qui sont alors parfaitement explicitées. Ainsi la théorie des ensembles n'est pas sortie toute casquée du cerveau d'un mathématicien génial (bien que Cantor le fût). Elle est la conséquence quasi inéluctable d'un développement de la science mathématique selon la méthode axiomatique ; elle est autant le fruit de la nécessité que du hasard.

L'ensemble $\mathbb{R}$ des nombres réels est, de par sa construction même, un des ensembles les plus riches (sinon le plus riche) des mathématiques. En effet, de multiples concepts y sont définis avec de nombreuses propriétés les reliant. En appliquant la méthode axiomatique, on dégagera trois structures fondamentales ; c'est-à-dire que l'on regroupera les concepts et leurs propriétés en très grands agrégats. Cette classification n'est pas arbitraire ; elle est le fruit de l'expérience acquise au cours du développement antérieur de la science mathématique. Le mathématicien entrevoit son intérêt scientifique en subodorant les généralisations auxquelles elle pourra conduire.

On distingue ainsi trois grands types de concepts définis sur $\mathbb{R}$ *et de nature mathématique différente :*

— des opérations (addition et multiplication) à l'aide desquelles il est possible de combiner certains nombres pour en obtenir d'autres ;

— une relation notée $\leqslant$ *qui permet de comparer deux nombres quelconques ;*

— une opération « valeur absolue » grâce à laquelle il est possible d'évaluer la proximité de deux nombres.

En opérant cette distinction, il est clair que l'on appauvrit sensiblement l'ensemble $\mathbb{R}$ *: ces trois types de concepts ne sont pas indépendants, car ils sont définis de manière constructive à partir d'un corps de concepts plus limité. Par exemple, la définition de la valeur absolue découle de celle de la relation* $\leqslant$ *puisque*

$$|x| = \begin{cases} +x & si \quad x \geqslant 0 \\ -x & si \quad x \leqslant 0. \end{cases}$$

Mais le mathématicien peut à loisir ignorer ces relations de dépendance et de compatibilité pour considérer chaque type de concepts de manière isolée et le soumettre à l'épreuve de la méthode axiomatique. Prenons l'addition, par exemple ; elle vérifie les propriétés évidentes :

(G1) $x + (y + z) = (x + y) + z$

(G2) $x + y = y + x$

(G3) il existe un nombre, 0, tel que $0 + x = x + 0 = x$

(G4) $x + (-x) = 0.$

Le mathématicien va induire de ces quatre propriétés une structure générale : la structure de groupe. *Il lui reste alors à déduire de manière rigoureuse toutes les propositions qui découlent de la définition par les propriétés G1, G2, G3, G4 pour voir se bâtir une nouvelle théorie, parfaitement indépendante de l'idée initiale qui avait consisté à isoler quelques concepts particuliers de* $\mathbb{R}$ *pour les généraliser. Cela étant fait, il pourra remarquer — à sa grande satisfaction — qu'il retrouve cette structure de groupe dans d'autres branches des mathématiques (groupes de transformations géométriques, par exemple) qui bénéficient du même coup de tous les résultats de la théorie élaborée. Il va sans dire que cela constitue une version très simpliste de la genèse de la théorie des groupes. Il a fallu beaucoup de temps et de mathématiciens pour qu'elle accède au rang de théorie !*

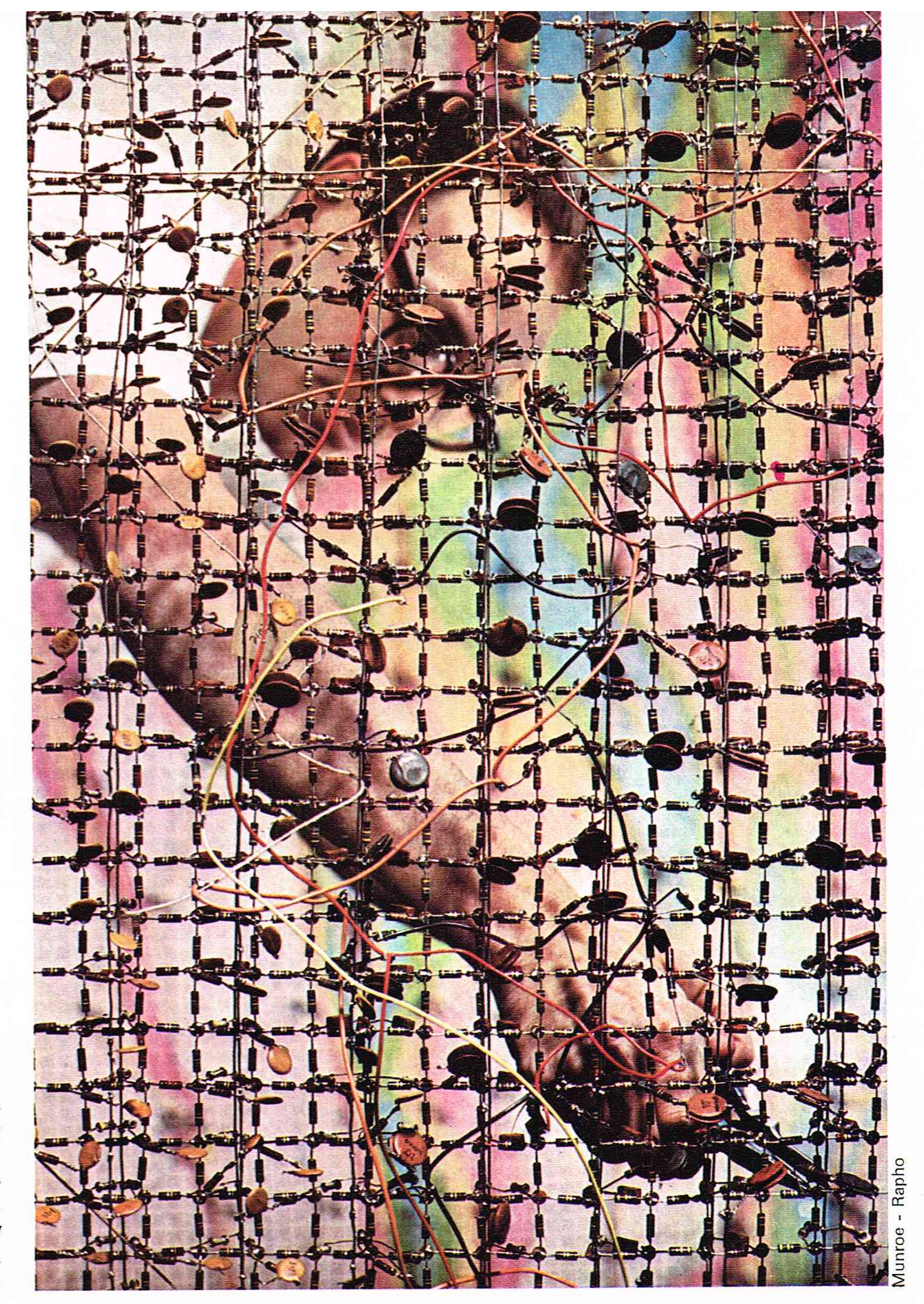
Structures en mathématiques

Il est malaisé de donner une vue d'ensemble de la science mathématique. On peut néanmoins affirmer sans crainte qu'elle ne se présente guère sous la forme stratifiée et officielle des programmes scolaires et universitaires, et qu'elle est aujourd'hui plus que jamais en pleine évolution, toujours en quête de sa propre unité.

Ainsi que nous venons de le voir à propos des nombres réels, trois grands types de structures doivent être distingués.

— Des structures algébriques *(groupe, anneau, corps, espace vectoriel...) qui traitent d'ensembles munis de lois de composition, grâce auxquelles on combine certains éléments pour en obtenir d'autres.*

— Des structures topologiques qui fournissent une formalisation abstraite des notions intuitives de voisinage, de limite, de continuité, etc.



▲ *L'omniprésente informatique est grande consommatrice de mathématiques.*

— Des structures d'ordre qui expriment la possibilité de comparer des éléments.

Structure algébrique ⟶ COMBINER
Structure topologique ⟶ APPROCHER
Structure d'ordre ⟶ COMPARER

Ces trois structures mères *constituent la charpente de l'édifice mathématique tel qu'il se présente aujourd'hui. Au-delà de ce premier noyau apparaissent ce que l'on pourrait appeler les* structures multiples, *qui se situent à la confluence de deux ou plusieurs structures mères. Ainsi, l'algèbre topologique est l'étude des opérations algébriques compatibles avec une structure topologique donnée, celles qui sont continues pour la topologie que l'on considère. L'analyse fonctionnelle est précisément de ce type-là, puisqu'elle développe l'étude des espaces vectoriels topologiques. De même, la topologie algébrique est l'étude des opérations s'effectuant sur certains ensembles de points définis par des propriétés topologiques (simplexe, cycle...). Son développement récent correspond à une tendance de la science mathématique à l'algébrisation de toutes ses branches, qui, issues de l'analyse, sont fondées sur des arguments et des procédés de raisonnement constructifs et non point déductifs.*

▶ *On devine l'angoisse de ceux qui ne comprennent pas, à dix ou douze ans (sans parler aussi de leurs aînés), qu'une droite puisse être « le supplément d'un hyperplan »...*

Mathématique et arbitraire

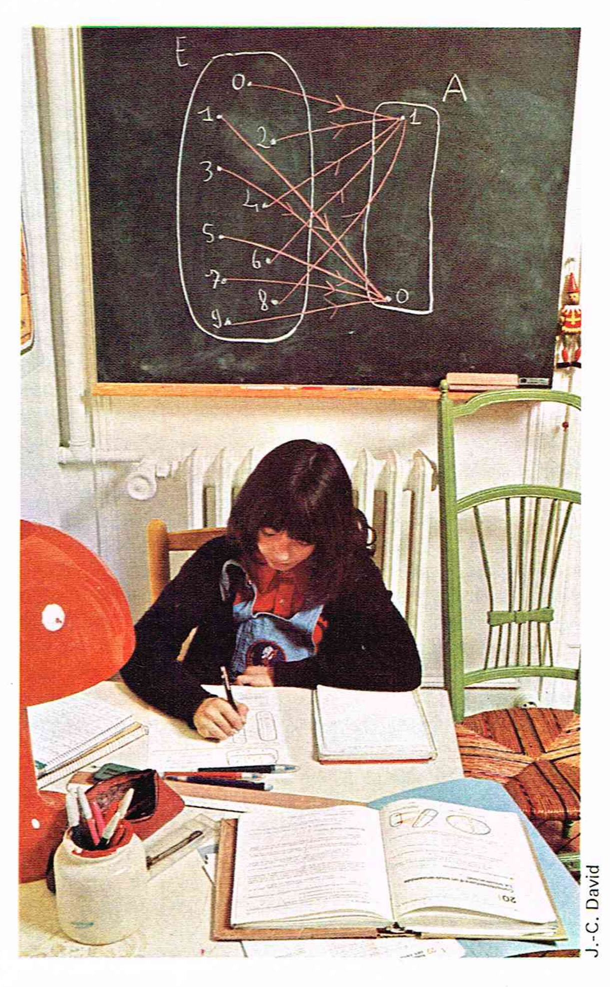
Nous avons montré que la démarche mathématique procède d'un balancement continuel induction ⇌ déduction. C'est l'assimilation abusive à la seule déduction qui, pour beaucoup, semble donner à la science mathématique un aspect « autoritaire » de science qui ne peut que dire la vérité. En réalité, il s'agit là d'un artifice qui consiste à oublier le problème de l'adéquation des axiomes de base à ce que l'on cherche à étudier : en écartant volontairement les cas qui ne vérifient pas les hypothèses qu'ils prennent pour base de leurs modèles, les économistes ou les physiciens font un choix sur lequel la discussion est souvent acharnée ; de même, en délimitant le cadre de l'étude par un système d'axiomes, les mathématiciens prennent une position qu'il importe de montrer sans ambiguïté dans son caractère non absolu.

Beaucoup d'exemples pourraient être donnés à ce propos, et le plus frappant est peut-être celui qui est fourni par la statistique, elle-même fondée sur le calcul des probabilités. Les magnifiques idées de Pascal nécessitèrent bien vite une formalisation claire de cet outil : il fallait définir la probabilité pour lui appliquer alors en toute rigueur les règles de la logique. Beaucoup (de Laplace à Poincaré et Hadamard...) se heurtèrent à la notion de hasard, et il semble bien que, depuis des siècles, ce mystérieux hasard soit ce qui sépare mathématiciens et philosophes réunis pour en discuter. Alors, malheureusement, on a parfois recours au formalisme le plus ardu pour définir un outil — la probabilité — que l'on ne cherche à justifier (si l'on suit l'ordonnancement scolaire ou universitaire) qu'après de longs et difficiles développements par la loi des grands nombres, loi qui, présentée au contraire dans son contexte concret, fait apparaître l'ambiguïté première du calcul des probabilités : elle oblige en effet à se poser de multiples questions, en particulier sur la répétition d'une expérience dans des conditions bien définies, et bien d'autres encore.

On ne peut donc pas assigner à la science mathématique cet aspect mystique, mais tellement stérile, de science « éthérée », dégagée de tout sectarisme, de tout subjectivisme, et qui ne se discute pas. Ce n'est pas le développement technique en lui-même d'une question (le calcul des probabilités, par exemple) qui peut être suspecté, mais le fondement de son développement.

De la pédagogie

L'autre aspect usuellement mis en évidence dans la caractérisation de la science mathématique est la nécessité de sa démarche logique et formelle, autrement dit de son procédé déductif. Le XIXe siècle a été celui d'une certaine révolution pour la pratique mathématique ; la rigueur de la déduction a été installée par l'axiomatique. C'est en ce sens qu'il faut considérer les mathématiques comme abstraites. Cependant, il ne faut pas assimiler rapidement le terme abstrait et le terme moderne, comme y invitent un grand nombre de clichés de notre vie de tous les jours. Tout d'abord, on ne devrait pas oublier que les « mathématiques modernes » ne sont en aucune façon une rupture avec les autres mathématiques (et l'on entend souvent par là, celles que l'on peut se représenter), et que l'on n'a fait que suivre



J.-C. David

l'évolution générale de la pensée humaine. Ensuite, que ce formalisme a déjà plus d'un siècle et ne répond donc sûrement plus à cette épithète, qui se trouve le plus souvent utilisée comme un épouvantail.

Il y a certainement un fond bien justifié à cette idée, car le souci de rigueur dans la méthode semble avoir conduit à de notables exagérations. Cela est d'autant plus clair après que l'on a bien mis en évidence la nécessité de la phase inductive dans la pratique mathématique. Nous reprendrons bien volontiers les termes de M.-D. Revuz, par lesquels il nous paraît cerner les problèmes de base que l'enseignement actuel des mathématiques n'a pas résolus : « Reprocher aux mathématiques d'être abstraites est une sottise : elles le sont par nature ; mais reprocher à un enseignement des mathématiques de ne pas montrer nettement d'où et comment les mathématiques ont été abstraites est légitime. » On pourrait résumer ces termes en disant que l'abstraction ne peut être prise pour une fin en soi, et que chaque concept mathématique abstrait est l'aboutissement d'une longue suite de développements sur des thèmes au départ plus familiers à notre compréhension. Autrement dit, il n'est possible de saisir avec fruit ces notions, fort peu « parlantes » en général, qu'à l'aide de leur histoire. Il y a une unité profonde de la science mathématique dans le temps et dans l'espace, et l'on devine l'angoisse de ceux qui ne comprennent pas, à dix ou douze ans (sans parler aussi de leurs aînés), qu'une droite puisse être « le supplémentaire d'un hyperplan », à moins qu'elle ne soit définie comme « un ensemble de points muni d'une bijection g

avec l'ensemble $\mathbb{R}$ telle que toute autre bijection f *de D sur* $\mathbb{R}$ *s'en déduise par translation :* f *(M)* = g *(M)* + a *»!*

C'est cette unité qui lui donne sa dynamique propre et lui permet de progresser, contrairement à une idée trop répandue — même par certain dictionnaire — selon laquelle la mathématique est figée, immuable, et n'a plus de secrets. Les mathématiques ont aussi des bases concrètes (et souvent, elles ont là la source de leurs plus grandes difficultés, ainsi l'idée de probabilité évoquée plus haut) et une histoire dont l'ignorance met une barrière à tout essai de compréhension ; harmonieusement liées, elles guident vers un exposé plus formel dégagé de toute ambiguïté et de tout ésotérisme.

Mathématique et histoire

Il n'y a pas plus de mathématiques modernes sans aucun lien avec l'édifice mathématique construit au fil des siècles qu'il n'y a de nouvelle physique relativiste contredisant la mécanique classique de Newton ; chacune est un modèle et correspond à une certaine échelle, les plus récentes tentant de représenter concurremment deux types de phénomènes tout en les unifiant. Ainsi, l'étude des figures géométriques que sont le cercle, l'ellipse, la parabole et l'hyperbole est-elle devenue l'étude « réunifiée » des sections planes du cône de révolution, le développement de l'algèbre linéaire et multilinéaire posant ensuite un cadre général aux courbes et surfaces du second degré, coniques et quadriques. On peut citer encore le cas de la droite, dont on a parlé plus haut, envisagée d'abord comme le plus court chemin d'un point à un autre (et donc définie par ces deux points), puis comme sous-espace vectoriel, enfin comme structure isomorphe à l'ensemble $\mathbb{R}$. *La notion de mesure fournit encore un exemple de réunification pour les notions de longueur, de surface, de volume, et même aussi de probabilité ; il est essentiel de ne montrer un développement théorique de l'intégration qu'au travers de ces idées, faute de quoi on ne joue qu'un jeu d'abstraction outrancière. On ne peut d'ailleurs comprendre où va et où peut aller l'esprit dans sa recherche de rigueur pour la construction mathématique qu'en évoquant les étapes antérieures et les nécessités qui ont poussé à leur formulation. C'est alors, et alors seulement qu'apparaîtront l'utilité du travail du mathématicien et sa place dans la démarche générale de la pensée humaine.*

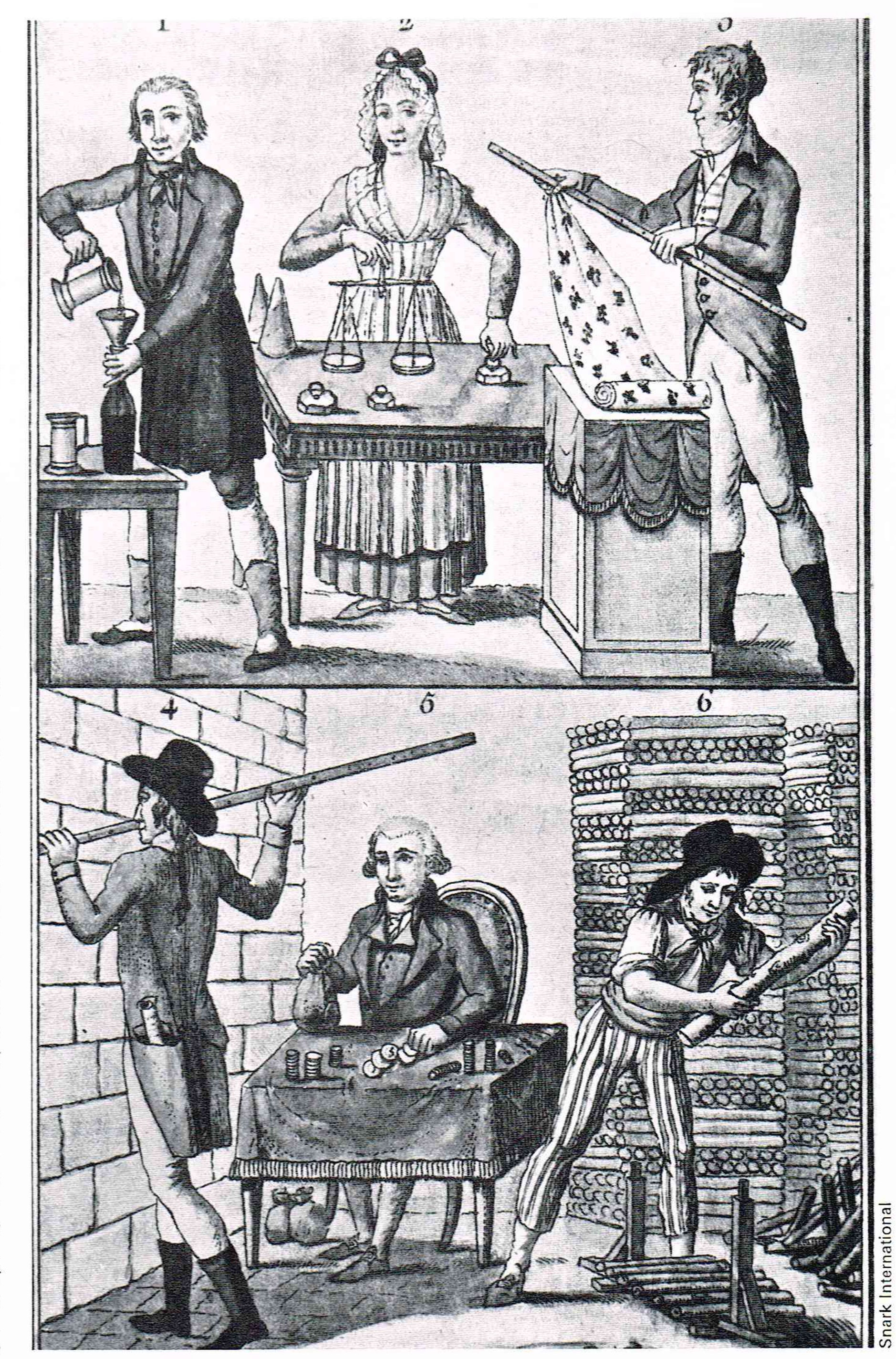
Il n'est pas possible de lire des mathématiques comme on lit un roman, sauf lorsqu'on ne cherche à tirer de sa lecture qu'un vague souvenir de titres et de sous-titres. Le but de la science mathématique ne peut être atteint que par des moyens (dont le moins important n'est pas le langage) dont la maîtrise ne peut s'acquérir qu'au terme d'une certaine pratique.

Nous avons donc pris le parti dans ce volume de ne pas rechercher un exposé didactique. La littérature mathématique est abondante et peut suffire à toute demande de ce type. Nous avons surtout cherché à donner ici un aperçu démystificateur sur la science mathématique. Nous avons donc pensé que le chapitre Histoire des mathématiques *devait être la pierre centrale de ce volume, et non pas une introduction à de multiples développements techniques. Il nous a donc semblé intéressant de placer cet aperçu historique en fin de volume, après les différentes matières traitées. Une première lecture des premiers chapitres pourra ainsi permettre de saisir plus complètement l'évolution des mathématiques.*

Il n'a pas été non plus dans notre intention de faire un exposé complet de chaque matière. Nous avons plutôt tenté de donner un aperçu simple des problèmes, tout en dépassant le stade de quelques considérations de salon. La bibliographie indiquée au terme de chaque chapitre donne, par contre, l'occasion à celui qui le désirerait de poursuivre ses idées sur la matière.

B. GOLDFARB

et J. THÉPOT



Snark International

▲ La notion de mesure fournit un exemple de réunification pour les notions de longueur, de surface, de volume, et même aussi de probabilité.

MATHÉMATIQUES ET PHILOSOPHIE

Au moment où la philosophie semble décriée, la philosophie mathématique n'est pas seulement vivante, mais douée de moyens techniques qui en assurent l'éternité. J. Hadamard remarquait justement, dans une préface (1926) aux *Fondements des mathématiques* de F. Gonseth, que c'est « un bien étrange phénomène, sans précédent dans l'histoire de la pensée », qu' « une science parvenue à l'état positif [soit] en train de faire marche arrière et de revenir à l'état métaphysique », ce qui est précisément le cas de la mathématique, réputée si simple, si ancienne, et si parfaite. Et de fait, il y a en mathématiques, depuis la fin du XIXe siècle, comme un retour à la philosophie, à une pensée profonde qui sonde l'être mathématique pour en formuler, enfin, la nature sans quitter le terrain de la rigueur, sans déroger aux principes qui ont établi la réputation de cette discipline depuis que la philosophie grecque l'a promue au rang de modèle de toute rigueur, de type idéal de tout discours, et en particulier du discours philosophique.

Les liens se sont tissés entre la mathématique et la philosophie depuis qu'il existe une mathématique rigoureuse, c'est-à-dire, hypothético-déductive. Ce n'est pas un hasard si l'idée d'*universalité* qui semblait caractériser les énoncés mathématiques de manière essentielle a présidé à la naissance des grands discours philosophiques. Maints philosophes auraient bien voulu qu'on pût dire aujourd'hui de certaines propositions philosophiques ce qu'on dit de la démonstration mathématique : que ce qui était une démonstration pour Euclide le demeure toujours pour nous. Malheureusement, le discours philosophique n'a jamais pu produire le système dont il rêvait. C'est pourquoi nous allons essayer d'abord de voir quelle fut l'intention première de ce discours, aujourd'hui périmé. Cela nous permettra ensuite de mieux mesurer la signification de la renaissance effective d'une certaine philosophie mathématique à partir du XIXe siècle. Mais auparavant, à quoi se réduit le projet philosophique traditionnel, contemporain de l'émancipation des mathématiques à l'égard des magies et des techniques, et devenu aujourd'hui sans importance, dans sa forme primitive, pour les mathématiques ensemblistes, axiomatisées, « appuyées » sur la logique mathématique contemporaine ?


Euclidis Megarensis accutissimi mathematici elementorum liber primus ex traditione Theonis Bartholameo Zāberto Vene. interpte icipit aue foelici.

Diffinitio prima.

Signum est cuius pars nulla.

Diffinitio. ii.

Linea uero longitudo illatabilis.

Diffinitio. iii.

Lineæ autem limites sunt signa.

Diffinitio. iiii.

Recta linea ē q̄ ex æq̄li sua iteriacet signa.

Diffinitio. v.

Superficies est quę longitudinem: latitudinemque tantum habet.

Diffinitio. vi.

Superficiei extrema sunt lineæ.

Diffinitio. vii.

Plana superficies est quæ ex æquali suas interiacet lineas.

Diffinitio. viii.

Planus angulus ē: duarū linearum in plano se se tangentium: & non in directo iacentium ad alterutram inclinatio.

Diffinitio. ix.

Quando autem quę angulū continent lineæ rectæ fuerint: recti lineus angulus nuncupatur.

Diffinitio. x.

Cum uero recta linea super rectam consistens lineam utrobiq; angulos æquales ad inuicem fecerit: rectus est uterq; æqualium angulorum: & quę supstat recta linea perpēdicularis uocitat sup q̄ steterit.

Diffinitio. xi.

Obtusus angulus maior est recto.

Diffinitio. xii.

Acutus uero minor est recto.

Diffinitio. xiii.

Terminus est quod cuiusq; finis est.

▼ *Page extraite des* **« Éléments »** *d'Euclide,* ***édition traduite en latin par Zamberto (Florence, Bibliothèque nationale).*** *Dans cette œuvre qui date de 300 avant J.-C., Euclide propose* ***la première*** **méthode axiomatique** *d'une théorie* ***mathématique, la géométrie.***

Pineider

Les rapports des mathématiques avec la philosophie

Le privilège de la « réflexion » ; le projet du rationalisme traditionnel

En fait, il semble que la *révélation* du discours démonstratif ait été, d'emblée et en même temps, la principale source de la philosophie comme discours rigoureux. Toutefois, le discours philosophique, qui aspirait à la rigueur, en transformant celle-ci en thème explicite propre tombant dans le champ de ses compétences, visait plus haut. Il voulait fonder d'abord cette rigueur mathématique qui avait le défaut d'être limitée aux mathématiques et, en tant que telle, était incapable de se fonder elle-même. C'est pourquoi l'on voit Platon, déjà, s'efforcer de développer une critique des mathématiques comme domaine scientifique spécifique, trop étroit pour le projet de critique et de fondation radicale qui anime la philosophie en tant que telle. Il fallait retrouver dans un discours plus primitif, dans une racine plus profonde, les raisons ultimes qui permettent de caractériser la rigueur mathématique par les propriétés qui lui reviennent en propre et en droit, et qui en font une activité intellectuelle indépendante de l'univers matériel et contraignante pour tout sujet. Or, ce discours, c'est la dialectique.

Le postulat général de toute la philosophie occidentale est par là suffisamment indiqué. Il ne dit rien d'autre que ceci : la possibilité, bien que chaque fois soumise à rude épreuve, d'un discours fondamental ou d'une démarche fondatrice qui livre la vérité ultime des techniques démonstratives. La théorie de la réminiscence est la première tentative pour expliquer le fait de l'indépendance de ces techniques envers toutes les contingences de nature subjective ou expérimentale, de leur apparence omnitemporelle, interprétée dans le sens d'objets idéaux, éternels, non soumis aux vicissitudes du temps, dans le sens d'une *mathésis universelle* et essentielle, comme l'on dira beaucoup plus tard. Or, toutes les grandes philosophies ultérieures, tous les rationalismes sont autant d'essais d'explication conçus pour les mêmes faits.

Ainsi, l'évidence cartésienne, par exemple. Elle est d'abord celle du *cogito,* du « je pense », manifestée au terme d'un doute méthodiquement mené comme seul foyer de résistance à ce doute, comme seul élément qui ne se laisserait réduire qu'au prix d'une réduction à l'absurde de l'ensemble de l'entreprise. La vérité mathématique elle-même ne résiste pas au doute : Descartes imagine un Malin Génie qui rendrait l'homme capable de douter de la vérité de $2 + 2 = 4$. Toutefois, je ne peux douter, dit-il, que *je* doute ; et que je puisse être trompé prouve encore que je pense, et donc que j'existe en tant que tel.

Si bien que, du point de vue du sujet connaissant, tout repose sur cette évidence, qui est un *intuitus* (intuition = vision) modèle, car le *cogito,* dont la connaissance n'est que l'établissement d'un point ferme dans le doute universel et radical, met fin à l'universalité du doute en instaurant une vérité sur laquelle s'articulera toute la chaîne des certitudes. Fournissant un échantillon de

l'évidence propre à une vérité indubitable, le *cogito,* dit un commentateur, permet de considérer comme telle tout ce qui possède une vérité comparable. Les énoncés mathématiques eux-mêmes ne constituent que des vérités relatives au *cogito,* parce que la vérité du *cogito* est encore plus simple que celle des vérités mathématiques. Celles-ci peuvent fonder les représentations sensibles ; celui-là fonde toute représentation possible. Il est condition nécessaire de tout ce que nous pensons ; car « il est impossible que nous puissions jamais penser à aucune chose que nous n'ayons en même temps l'idée de notre âme », que le *cogito* n'accompagne notre pensée.

Est-ce autre chose qu'entreprend Kant ? Certes, si l'on considère la différence d'horizon, due essentiellement à l'importance, aux yeux de Kant, de la physique newtonienne ; en second lieu, à une différence de langage, due elle aussi à la constitution, au temps de Kant, d'une science rationnelle de la nature ; en troisième lieu enfin, au fait que sa conception ne jaillit pas d'innovations techniques, mais d'une interprétation des propositions élémentaires de la mathématique traditionnelle. Soit la géométrie euclidienne, par exemple. On ne peut dire qu'elle se fonde sur l'expérience. Ce serait attenter à son universalité. Or, elle se fonde sur l'idée d'espace. Celle-ci ne pouvant être un donné empirique, elle ne peut être un pur concept. Si la notion de concept est distincte ici de celle de concept *abstrait,* c'est-à-dire extrait de l'expérience, elle ne repose pas non plus sur un concept *pur,* c'est-à-dire qu'en dernier ressort, elle est irréductible à la pure logique. N'étant ni purement logique, ni expérimentale, on lui crée une forme nouvelle d'état civil : le concept d'*intuition pure,* par lequel Kant limite le pouvoir de l'entendement, disons, de logique pure, à la forme de la sensibilité. Cette forme, celle de l'espace en l'occurrence, est irréductible à un concept, comme l'illustre l'exemple de l'égalité indirecte des figures, qui, ainsi que les mains, ne sont superposables que par retournement dans l'espace, c'est-à-dire que la symétrie ne conserve certaines propriétés de la figure qu'en inversant celle-ci, phénomène dont l'analyse logique ne peut nullement rendre compte, ce qui est vrai tant que l'on ne dispose pas du groupe des transformations qui conservent les angles en grandeur tout en changeant leur signe. Il en résulte pourtant que les jugements mathématiques sont *synthétiques* (irréductibles à la logique) et *purs* (non d'origine expérimentale). Or, non seulement c'est le philosophe critique qui met au jour cette vérité de nature d'ailleurs philosophique, mais de plus l'idée de synthèse rationnelle est au principe de toute science. Et nulle science n'échappe à la fondation de mode désormais *critique*.

Après Kant, Husserl retrouve une inspiration plus cartésienne, mais inscrite dans le projet général et rationaliste tout à fait dans la lignée de Platon, de Descartes et de Kant. Husserl entreprend, dès les *Recherches logiques,* de montrer que la mathématique et la logique constituent un domaine qui ne peut se confondre avec les conditions subjectives qui permettent d'en prendre connaissance. Par là, Husserl suit l'enseignement de son grand contemporain, G. Frege, pour lequel les mathématiques (en vérité l'arithmétique), étant un prolongement de la logique, ne peuvent être élucidées par l'étude de leur genèse psychologique. Mais Husserl ne suit pas Frege jusqu'au bout ; il réeffectue la philosophie cartésienne dans la mesure où il analyse la constitution de la « conscience » comme région privilégiée et fondatrice de toute expérience, systématiquement explorée dans les *Méditations cartésiennes,* qui s'offrent comme un *recommencement* radical de la philosophie, comme (d'après J. T. Desanti) l'effort le plus têtu de Husserl, et peut-être de toute la tradition rationaliste occidentale, pour ressaisir et repenser le projet d'une fondation des mathématiques, et par suite des sciences, ailleurs qu'en elles-mêmes, sur une évidence ultime, apodictique (universelle et nécessaire), en tant qu'« intuition » (ou évidence) pure, et pour cette raison susceptible de servir de support à des jugements apodictiques.

Tous les cas considérés ici — on a laissé de côté la tradition empiriste qui est, sous sa forme ancienne et classique, condamnée dès l'origine à la stérilité sur ce point, puisqu'elle professe que les concepts mathématiques résultent d'une genèse empirique jamais démontrée, ou prouvée, ou construite — procèdent à un essai de subordination du savoir mathématique, à son enracinement dans un sol que met à nu une analyse soit dialectique, soit métaphysique, soit transcendantale. Bien que rien de commun ne relie la nécessité de déployer le contenu lui-même du savoir à celle de définir *a priori* ses conditions de possibilité, on trouve à la racine de tous ces projets l'idée que le savoir mathématique a des secrets que seule peut livrer la philosophie. Bref, qu'il ne vaut que dans la mesure où il peut être « intériorisé » par un discours distinct de lui, et qui le dépasse.

Le rôle actif de la philosophie : l'idée de discours démonstratif

Mais reconnaître la tendance et le caractère « intériorisateurs » et « assimilateurs » de l'investigation philosophique traditionnelle ne devrait pas occulter le fait que la forme de vérité visée dans cette investigation a, en fait, des rapports plus complexes avec la mathématique comme théorie, et l'idée de théorie comme propriété permanente du rationnel et du scientifique. A cet égard, il y a si peu de discontinuité et de contraste entre l'exigence philosophique et l'exigence mathématique qu'à l'origine, c'est la mathématique qui a, pour ainsi dire, assimilé et intériorisé le projet philosophique.

L'Antiquité grecque

En effet, il est au moins une chose bien certaine, c'est que la figure de la mathématique grecque, où l'idée de mathématique s'est identifiée à celle de démonstration, est fondamentalement différente de la mathématique d'origine égyptienne et babylonienne. Ce qui fait cette différence, c'est sans doute, non le caractère *rigoureux,*

▲ ***La* Philosophie *entourée d'*Aristote, *de* Platon, *de* Socrate *et de* Sénèque *: planche ancienne tirée du* Panégyrique de B. Visconti *de B. de Bologna (XIVe siècle).***

Giraudon

dans la mesure où une technique pure peut bien être rigoureuse, mais le caractère méthodiquement discursif, l'idée d'un système hypothético-déductif, c'est-à-dire fondé sur des hypothèses de départ explicites, dites axiomes ou postulats, dont on examine les conséquences logiques. D'où la mise en œuvre d'une technique démonstrative tellement précise et rigoureuse qu'elle reste encore pour nous un modèle de rigueur. Une technique : Euclide procède toujours à partir d'un énoncé qui est donné comme hypothèse, qu'il explique par un cas « concret » pour ensuite le prouver, le *démontrer*. La démonstration est ici le moment le plus important, le plus spécifique des textes mathématiques grecs. Or, elle est reliée à une tendance profondément hostile à l'expérience et à l'intuition, exprimée avec force par Platon qui veut que l'énoncé arithmétique porte « sur les nombres en tant que nombres, sans admettre d'aucune manière qu'on fasse ce discours en lui proposant des nombres pourvus d'un corps visible et tangible » (*République*, 525 d). Il y a donc une science des nombres à partir du moment où l'on tourne le dos à l'arithmologie figurée. Il y a une science des nombres dans la mesure où **le** nombre ne se confond pas avec **les** nombres concrets. D'ailleurs, cela vaut exactement pour la géométrie : elle nous intéresse, non parce qu'elle sert « pour l'établissement d'un camp, pour le siège d'une place, pour la concentration ou le déploiement d'une armée, ou pour tout ce qu'encore un plan de bataille, un ordre de marche comportent en eux-mêmes de figures », mais parce qu'elle « force à contempler la *réalité* » (*République*, 526 d-e), dans la mesure où celle-ci n'est ni l'utilité, ni la technique des figures « visibles et tangibles », qui ne peuvent en aucun cas fournir le point de départ à la géométrie comme science. Pour cette raison, Euclide, fidèle à la tradition exprimée par Platon, manifeste constamment le souci d'affranchir autant que possible ses définitions et ses démonstrations de toute représentation tactile ou visuelle. Il ne considère pas telle ou telle figure, mais toutes les figures de tel type : *Dans* **tout** *parallélogramme,* dira-t-il, *les compléments des parallélogrammes situés de part et d'autre de la diagonale sont équivalents entre eux.* Ou encore : **Toute** *pyramide à base triangulaire peut être divisée en deux pyramides à base triangulaire équivalentes et semblables (c'est-à-dire égales) entre elles et semblables à la pyramide entière, et en deux prismes équivalents, et la somme des deux prismes est supérieure à la moitié de la pyramide entière.* Si l'on considère **tout** parallélépipède, **toute** pyramide, c'est que ce n'est pas **ce** parallélépipède ni **cette** pyramide qui donnent la preuve exigée par la géométrie. C'est donc en persévérant dans la voie de l'épuration logique, de la rigueur démonstrative, en réalisant une intention qui anime l'écriture mathématique d'Euclide, que Leibniz remarquera plus explicitement encore que : « La force de la démonstration est indépendante de la figure tracée, qui n'est là que pour faciliter l'intelligence de ce qu'on veut dire et fixer l'attention ; ce sont des propositions *universelles,* c'est-à-dire les définitions, les axiomes et les théorèmes déjà démontrés qui font le raisonnement et le contiendraient quand la figure n'y serait pas. » C'est pourquoi, déjà, J. Schenbelius et C. Herlinus ont, dès le XVIe siècle, l'un donné des figures sans leurs lettres, l'autre réduit les mêmes démonstrations en syllogismes et prosyllogismes. Incontestablement, logique et mathématique sont liées chez Euclide, bien que la logique ne se révèle que par l'exigence, encore non satisfaite, d'un discours pur, que ne justifient ni l'expérience visuelle, ni le souci de formules utiles permettant l'évaluation d'une aire ou d'un volume, ni le recours à l'intuition.

Une théorie extrêmement séduisante a été développée par l'historien A. Szabó pour expliquer cet état de fait : elle consiste à mettre en rapport l'anti-empirisme et l'hostilité à l'intuition avec la découverte du raisonnement indirect. Cette thèse a le mérite de souligner que le raisonnement indirect n'est ni intuitif ni empirique, et qu'il ne recourt ni au « visible », ni au « tangible », ni même au calculable. Mais elle suggère également que, si les premiers mathématiciens ont recouru à cette technique, c'est parce que les *monstrations,* empiriques et intuitives, ne les satisfaisaient plus. Ainsi, jamais l'attitude spontanée ou commune n'aurait pu concevoir l'échec des tentatives empiriques en vue de déterminer une mesure commune entre la diagonale du carré et son côté. Il fallait s'apercevoir que la nécessité d'admettre l'existence de grandeurs géométriques incommensurables entre elles est absolument incontournable, si incontournable que, derechef, on doit produire le concept d'incommensurabilité, qui ne vaut que dans la mesure où « côté », « diagonale », « carré » ont acquis le statut d'objets mathématiques, c'est-à-dire une « réalité » du même type que celle de l' « objet » qui s'est révélé incommensurable. La technique de preuve induit un *nouvel objet* dont la nature exige que les anciens objets soient repensés conformément à la nature qu'ils révèlent. C'est en cela que consiste ce qu'on a appelé « révolution » ou « miracle » grec.



Bulloz



Palais de la Découverte

▶ *A gauche,* ***planche ancienne du XVe siècle illustrant les rapports des mathématiques avec la philosophie.*** *A droite,* ***le philosophe et mathématicien*** **Bernhard Bolzano** *(1781-1848).*

Mais cet abandon de l'expérience dans des disciplines qui tiraient leur prestige de leur utilité n'est-il pas étonnant ? Certes. Soulignons bien, avant de donner une explication de ce phénomène, la place de choix que le raisonnement indirect occupe dans les *Éléments* d'Euclide. Il y a tout lieu de penser qu'au temps de Platon, il fut considéré comme la preuve mathématique par excellence. Ainsi, c'est après avoir, grâce à un raisonnement indirect, mis Théétète dans l'embarras, que Socrate semble réclamer autre chose que des arguments vraisemblables « dont l'usage en géométrie ferait que Théodore (maître de Théétète et géomètre lui-même) ne vaudrait pas même un seul liard s'il y recourait ». Comment alors expliquer ce fait ? Et cet autre qui l'accompagne : l'orientation anti-empirique ? Il est difficile de croire que, spontanément, une mathématique empiriste se soit muée en théorie d'idéalités mathématiques. Il est difficile, lorsqu'il s'agit de la naissance des mathématiques en Grèce, et donc de l'idée de démonstration, de faire abstraction de l'influence de la doctrine éléatique.

Les philosophes d'Élée se caractérisent, en effet, par le mépris qu'ils professèrent pour tout ce qui est empirique. La vérité, selon Parménide, ne peut être saisie que par la pensée (λόγῳ). Et c'est Parménide qui, le premier, applique les principes logiques de la non-contradiction et du tiers exclu pour établir que l'être ne peut être soumis à la génération. Car, ou bien il se serait produit à partir d'un *être,* ou bien d'un non-être. Dans le premier cas, il y aurait eu un être avant l'être ; dans le second, un être qui aurait été un non-être. Il en résulte, par un raisonnement typique et spécifique des Éléates, que l'être n'est pas soumis au devenir. Après Parménide, Zénon, auquel on attribue les fameux paradoxes, est un des plus illustres représentants de cette école de pensée. C'est pourquoi Aristote le considère comme le créateur de la dialectique, et l'on peut considérer la dialectique comme la créatrice des mathématiques, comme l'institutrice de l'usage des hypothèses, et, pour tout dire, de la méthode hypothético-déductive. Si bien qu'il apparaît qu'à l'origine, la mathématique n'était qu'une branche développée de la dialectique. C'est ce que peut montrer une analyse philosophique du vocabulaire des mathématiques grecques, à laquelle a procédé A. Szabó, et que nous ne pouvons ici examiner dans le détail.

Ce qui est certain, c'est que le souci, typiquement grec, de logique dans une discipline comme les mathématiques a conduit à la création d'une logique, c'est-à-dire d'un exemple de discipline démonstrative non mathématique *(cf.* la partie historique du chapitre sur *la Logique).* En effet, « la logique est aussi susceptible de démonstrations que la géométrie, dira plus tard Leibniz, et l'on peut dire que la logique des géomètres ou les manières d'argumenter qu'Euclide a expliquées et établies en parlant des propositions sont une extension ou promotion particulière de la logique générale ». L'œuvre logique d'Aristote, inventeur de la logique formelle, apparaît à Kant, selon une formule devenue célèbre, si parfaite que, depuis son fondateur, elle n'aurait fait « aucun pas en avant et que, par conséquent, selon toute apparence, elle semble close et achevée ». Ce jugement est certes aujourd'hui dépassé. Mais cela ne peut faire oublier qu'Aristote a précisé l'idée de déduction en énumérant quelques règles valides, et qu'il a doté en outre la logique d'une sorte d'axiomatisation. Qui plus est, il semble n'avoir médité sur les règles du raisonnement qu'en songeant à établir les principes de la discussion. Ainsi, bien que le terme *dialectique* ait chez lui un sens différent de celui de Zénon ou de Platon, c'est en vue de réglementer une dialectique envahissante qu'est née la logique d'Aristote.

Le renouveau du XIXe siècle

Il ne serait pas exagéré de dire que la renaissance de cet esprit *dialectique* au XIXe siècle n'a pas été étrangère aux retrouvailles entre logique et mathématique. Mais cet esprit dialectique n'a rien à voir avec la dialectique hégélienne ; il se nourrit plutôt de l'exigence de faire revivre la philosophie dans l'horizon même de la mathématique. Ou, si l'on veut, c'est dans la mathématique elle-même que se ressent le besoin d'une réflexion logique et philosophique.

Pour comprendre cette situation, il nous faut d'abord rappeler que toute l'originalité de la mathématique moderne consiste à avoir introduit l'idée de *variation.* Il n'en pouvait être question dans l'Antiquité grecque : la critique que Platon adresse aux géomètres de son temps exprime une exigence respectée ; elle consiste à dénoncer l'inadéquation entre la nature de la science et le langage employé par ceux qui la manient. « Leur langage, dit Platon, dans *la République* (527 a), est, je pense, tout à fait risible et sent la servilité ; car, tenant un langage qui est celui de gens qui pratiquent une action et dont la pratique est le but, ils parlent de « carrer », de « tendre le long de », de « poser au plus de » [...] alors que cette étude est tout entière une occupation dont *la connaissance* est le but. » Mais la mathématique moderne ne renie pas ce que les mathématiciens grecs connaissaient sans lui reconnaître de statut théorique. Au cours, donc, du XVIIe siècle, cette idée de variation s'impose. Elle conquiert un statut mathématique par la notion de *fonction* « qui s'introduit, remarque Bourbaki, à juste titre, et se précise d'une foule de manières au cours du XVIIe siècle », car, ajoute-t-il, « toute cinématique repose sur une idée intuitive, et, en quelque sorte, expérimentale, de quantités variables, avec le temps, c'est-à-dire de fonctions du temps ».

Le temps dans les mathématiques ? Il y a là de quoi scandaliser les amoureux de la rigueur, surtout depuis que la notion de fonction devait être, au XIXe siècle, élargie de manière que, par l'expression $y = f(x)$, on ne comprenne rien d'autre que le pur concept de l'opération « f » faisant correspondre au *système* des objets « x » le *système* des objets $f(x)$. Dirichlet disait, dès 1837, qu'il n'est pas nécessaire que y dépende de x *selon la même loi ;* il n'est même pas nécessaire d'avoir une dépendance exprimable par des opérations algébriques. Épurer les mathématiques, « omnitemporelles » par excellence, de leur « chute » dans le temps, devenait ainsi une entreprise mathématiquement pensable, aussi bien que logiquement nécessaire. Dans cette voie s'était déjà avancé B. Bolzano qui déclare péremptoirement dans ses fameux *Paradoxes de l'infini* (1850), lorsqu'il critique (§ 12) la notion d'infini telle qu'elle est exposée par l'article *Infini* du dictionnaire de Klugel, que même nos mathématiciens éminents, comme Cauchy, se sont trompés en pensant l'infini « comme une grandeur variable dont la *valeur* croît *indéfiniment* et, par suite, devient plus grande que toute *grandeur donnée* aussi grande qu'elle soit ». La limite de cette croissance infinie serait la *grandeur infiniment grande*... Ce qui est erroné dans cette définition éclate à partir du fait que ce que les mathématiciens appellent une *grandeur variable* n'est pas à proprement parler une *grandeur,* mais le pur et simple concept, la pure et simple *représentation* d'une grandeur qui n'est pas unique mais qui comporte un ensemble infini de grandeurs, de valeurs distinctes ». Évidemment, cette thèse de Bolzano n'acquiert toute sa portée que si le domaine infini des nombres réels est correctement construit. Cependant, grâce à la mise au jour du soubassement ensembliste de l'analyse classique, il aboutit à éliminer de l'idée de fonction toute idée de *variable au sens intuitif,* liée généralement à la notion de temps.

Or, Bolzano est à la fois un *philosophe,* qui s'est intéressé aux mathématiques parce qu'on s'y donnait la peine de *démontrer* des choses *évidentes,* et un *mathématicien.* Il a voulu allier les exigences de la philosophie et des mathématiques, et c'est dans cette association qu'il a pu réveiller l'intérêt pour la logique chez les mathématiciens, et l'intérêt pour les mathématiques chez les logiciens, à un moment où les logiciens se recrutaient essentiellement parmi les philosophes.

Toutefois, le plus curieux dans l'histoire, c'est que ce moment « bolzanien » de la mathématique moderne reproduit les éléments essentiels de la mathématique ancienne. En effet, Bolzano est philosophiquement platonicien. Il croit, comme Weierstrass plus tard, à l'existence en soi des êtres mathématiques ; il est anti-empiriste parce que le souci logique le pousse à mettre en lumière le fondement des mathématiques qu'il découvre lié à des notions que l'expérience ne peut fournir : par exemple, l'infini actuel ; enfin, il utilise le raisonnement indirect, dans l'idée que ce qui *n'est pas contradictoire est vrai* (voir dans le chapitre *Logique* les précisions qu'a apportées la logique moderne sur ces deux notions) et que de la non-contradiction à l'existence, la conséquence est bonne.

C'est encore plus résolument que le fondateur de la logique et de la philosophie contemporaines, G. Frege, s'est engagé sur cette voie de l'association philosophico-mathématique, mariage dont est née la logique mathématique pour laquelle Boole, d'une certaine façon, n'a pas fait plus que Leibniz. Or, l'originalité de Frege est d'avoir continué l'analyse et poursuivi l'élargissement de la notion de fonction jusqu'à créer une *jonction réelle* entre mathématique et logique, qui lui a permis, dans une entreprise partiellement manquée, de reconstruire logiquement la mathématique ou ce qui, de celle-ci, se réduisait à l'arithmétique des entiers naturels, ce qui n'était pas encore, du temps de Frege « rigoureusement » réalisé pour la géométrie. Or, Frege est, comme Bolzano, platonicien : « Que trois tombe sous le concept de nombre premier est une vérité objective ; ce n'est pas une idée que je trouve en moi ; c'est un fait toujours et objectivement valable, que nous dormions ou veillions, que nous le pensions ou pas. » Mais il est, en outre, comme on sait, logiciste, c'est-à-dire que, pour lui, les ressources logiques sont nécessaires et suffisantes pour penser l'arithmétique et tout ce qui peut se construire à partir d'elle, c'est-à-dire beaucoup de choses.

Il ressort de ces indications que la mathématique a partie liée avec la philosophie au moins dans ses moments cruciaux : celui où elle naît, où naît l'idée de démonstration, et celui où elle renaît pour réinvestir cette ancienne idée de la rigueur dans les nouveaux domaines conquis depuis le XVIIe siècle, c'est-à-dire essentiellement le calcul infinitésimal, et plus généralement l'analyse. Nous avons soigneusement laissé de côté le fait, somme toute contingent, qu'au XVIIe siècle, certains philosophes étaient mathématiciens : Descartes et Leibniz l'attestent, dont les philosophies mathématiques respectives dépendent de la mathématique que chacun avait en vue. Mais les mathématiques se sont développées après eux avec autant de rapidité que de désordre.

C'est après cette floraison technique que reparut l'exigence de tout remettre en ordre, d'entreprendre la révision des principes, de dépasser les vues directement reliées aux applications, de substituer, comme disait Lejeune-Dirichlet, les idées au calcul. Cette substitution, ce retour au concept n'allaient cependant pas s'accomplir sans difficultés. Ils sont à l'origine de la problématique des fondements.

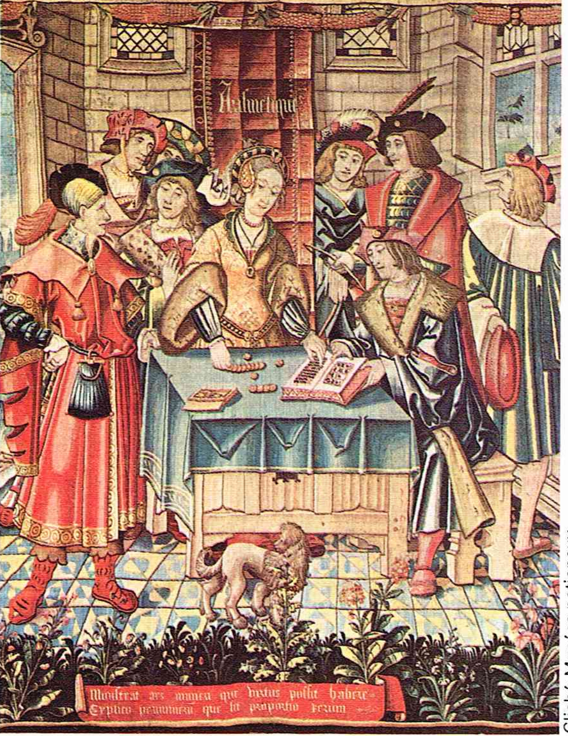
La fondation de la nouvelle philosophie

Si l'on considère les discussions du début de ce siècle sur les mathématiques, on s'aperçoit, à les comparer avec les philosophies mathématiques de Descartes et de Kant, qu'elles ont pour ainsi dire élargi l'horizon de la réflexion philosophico-mathématique. L'idée de mathématique a pris comme un nouveau départ, et les différentes élaborations qu'elle a suscitées apparaissent comme autant de manifestations décisives des ressorts de la philosophie mathématique contemporaine dans la mesure où elle a rendu la philosophie à son lieu, aux mathématiques, à un lieu dont le seul problème est de savoir s'il se maintient et comment, mais qui incontestablement est élaboré et produit par **G. Frege**.

D'abord, ce n'est donc pas un hasard si G. Frege constate, dans ses *Fondements de l'arithmétique* (1884), que son « exposé a pris un tour plus philosophique qu'il ne semblera à beaucoup de mathématiciens... une recherche fondamentale sur le concept de nombre ne peut manquer d'être marquée de philosophie. La tâche est commune aux mathématiques et à la philosophie ». S'il s'est engagé dans cette voie, ce n'est pas parce qu'il a dû professer, comme nous l'avons déjà indiqué, que les nombres sont des êtres mathématiques existant indépendamment de la pensée humaine et constituant un monde qui n'a pas besoin d'être fondé par autre chose que la logique pour pouvoir être compris ; c'est également pour deux autres raisons, plus essentielles.

La première concerne la signification de l'œuvre *logique* de Frege, inaugurée par la *Begriffsschrift* (1879) ou idéographie, l'ouvrage de logique le plus original et le plus profond depuis Aristote. Dans l'esprit de Frege, cette idéographie conduisait naturellement à élaborer une théorie générale des suites (Reihenlehre) dans le dessein de montrer que le principe de récurrence, l'un des axiomes de l'arithmétique, peut être élucidé de manière purement logique. Ce faisant, il montrait, contre la logique alors dominante (on est en 1879), celle de S. Mill, que l'empirisme n'a pas son mot à dire dans ce domaine. Qu'il est oiseux, voire incongru, de chercher dans les théories empiriques de l'induction une explication à l'induction complète. Du coup, et pour la première fois de manière convaincante, le rapport entre les mathématiques et les sciences de la nature se trouve irrémédiablement rompu, et rend impossible toute référence à la nature, à la physique ou à la subjectivité, auxquelles on recourait quand l'essentiel de la philosophie consistait, de Descartes à Kant, à expliquer le monde de l'expérience.

La seconde se rapporte à la perspective philosophique propre à Frege. Celle-ci est originale au sens fort. Elle constitue l'humus et l'horizon de la philosophie post-frégéenne, dans la mesure où elle a contraint le néo-kantisme, la phénoménologie naissante, la pensée analytique dans ses versions linguistique et épistémologique, à tenir compte de ses interrogations. Or, pour Frege, comme l'a souligné son meilleur commentateur, M. Dummet, la première tâche de la philosophie, ou plus exactement de l'enquête philosophique, c'est l'analyse du *sens*. Mais qu'est-ce que le sens ? R. Thom, proche d'une inspiration frégéenne, nous dit : « Tout mathématicien doté de tant soit peu d'honnêteté intellectuelle reconnaîtra que, dans chacune de ses démonstrations, il est capable d'attacher *un sens* à chacun des symboles qu'il manifeste : en cela, il diffère du physicien théoricien, qui, très fréquemment, n'hésite pas à se confier magiquement aux vertus du formalisme aveugle dans l'espoir — souvent déçu — que les lumières de la fin dissiperont les ténèbres du commencement. » Certes, cela veut dire que les mathématiques ont un *contenu propre,* et cela dit ce que Frege voulait dire. Mais la portée de l'analyse du sens qu'il institua va plus loin : il s'agit de montrer que toute thèse philosophique est dénuée de sens tant que l'on ne s'est pas entendu sur l'analyse du sens de certaines expressions. Il faut avant tout élucider le sens des expressions en jeu dans un problème pour pouvoir le poser, pour pouvoir se demander si ce qu'on dit est vrai et pour quelles raisons cela le serait. Autrement, il se peut qu'interviennent non seulement des différences de style dans l'analyse, mais des malentendus sur ce dont il est question, sur la chose analysée elle-même. La philosophie est alors œuvre d'élucidation, d'explicitation du langage. Par suite, il nous faut savoir comment travaille le langage, et donc établir la manière correcte de son fonctionnement. Autrement dit, il nous faut construire une théorie logique, et une théorie logique qui soit conforme à la *philosophie du sens,* une théorie



Cliché Musées nationaux

▶ ***Tapisserie française du XVIe siècle représentant « Madame Arithmétique » prodiguant ses enseignements à un groupe de jeunes nobles (musée de Cluny, Paris).***

de la définition et de l'inférence valide qui satisfasse l'exigence de cette philosophie, une théorie de la vérité qui rende compte de la manière dont elle se rapporte à — et utilise — les procédures du langage. Donc quelque chose, en fin de compte, d'encore plus profond que ce qu'on exige pour que le discours mathématique usuel soit entendu, c'est-à-dire de plus profond qu'une sémantique ensembliste.

Il y a comme l'ouverture à une ontologie précise dont on ne peut pas dire qu'aujourd'hui les problèmes soient clos. On a comme l'écho du grand désir de Platon de parvenir « à un principe qui se suffise » (*Phédon,* 101 e). Mais l'écho seulement. Car il ne s'agit pas de produire une dialectique supérieure au discours qu'elle analyse, mais d'asseoir, analytiquement, pas à pas, le discours analysé sur le langage qu'effectivement il implique.

Le problème des fondements des mathématiques

Mais la perspective, enfin mise au jour par Frege, n'était pas simple. Il ne s'agissait pas seulement de spéculer à nouveaux frais à partir de cette interrogation originaire, mais d'établir précisément l'arithmétique sur autre chose que le hasard des signes, sur des lois logiques vraies. Or, au bout de cette entreprise enfin réalisée dans les monumentales *Grundgesetze der Arithmetik,* s'est découverte une antinomie communiquée à Frege par une lettre célèbre de B. Russell.

Les paradoxes

Découverte en même temps, et sous une autre forme, par Cantor (1899) et par Zermelo, l'antinomie de Russell fut suivie par la mise au jour d'autres paradoxes, parmi lesquels on distingue aujourd'hui ceux qui sont de nature *logique* et ceux qui sont de nature *sémantique.*

A titre d'exemple, disons brièvement en quoi consiste le paradoxe de Russell (1902) : on appelle ensemble une collection arbitraire d'objets qui sont dits appartenir à cet ensemble. Les ensembles peuvent appartenir à, ou, en d'autres termes, être éléments d'autres ensembles. Mais un ensemble peut-il s'appartenir à lui-même ? Non, sans doute, car l'ensemble des chats n'est pas un chat. Pourtant, il existe des ensembles qui peuvent s'appartenir à eux-mêmes : l'ensemble de tous les ensembles. Si l'on considère E comme l'ensemble des ensembles F tel que F n'est pas un ensemble de lui-même, il en résulte que E est un élément de E si et seulement si E n'est pas un élément de E.

La nature des paradoxes est soit purement mathématique, exigeant alors une reconstruction de la théorie des ensembles jusqu'alors disponible dans les premiers travaux de Cantor, soit plutôt épistémologique, exigeant une logique qui précise le sens de la notion de vérité, de désignation, d'expressibilité, etc. C'est précisément ce qui montre que la philosophie mathématique va s'engager nécessairement sur deux voies appelées à se conjoindre et à se consolider mutuellement : mathématique (ensembliste) et logique.

Le logicisme

Le logicisme, c'est la reprise de la thèse de Frege. Mais, dans la version de **B. Russell,** cette thèse professe que *toutes* les mathématiques se réduisent à la logique ; et cela est affirmé dans l'horizon nouveau constitué par la nécessité de reconstruire l'édifice mathématique en évitant les antinomies. En ce sens, le logicisme est un réductionnisme. Il confond en une les deux solutions exigées par la théorie des ensembles : la solution mathématique et la solution sémantique. Logique et mathématiques sont donc une seule et même chose : la logique mathématique ; ce qui met en relief le caractère mathématique de la nouvelle logique, qu'on peut distinguer de la mathématique pure, comme le montre l'œuvre gigantesque des *Principia mathematica* (1910-1913), dont les trois volumes, rédigés en collaboration avec Whitehead, ont été la Bible des logiciens mathématiciens jusqu'aux premiers manuels de logique dus aux disciples de Hilbert : celui de Hilbert-Ackermann (1928) et celui, plus important, de Hilbert-Bernays (1934-1939).

Mais l'intérêt de la position logiciste ne vient pas tant de son caractère plus philosophique que technique, que du fait que ce caractère est lui-même à l'origine d'une investigation systématique sur les paradoxes, qui a conduit, comme l'a remarqué Gödel, à la mise en lumière du caractère contradictoire de nos intuitions logiques, et à la constitution d'une « doctrine des types » (B. Russell *in Principles of Mathematics,* 1903, App. B).

La distinction entre les types logiques semble, pour B. Russell, fournir la clef du mystère. Nous pouvons remarquer tout d'abord que cette distinction introduit des différences entre les divers niveaux des concepts. Par là, elle met en œuvre une analyse du sens — ou de la signification — conformément à l'exigence de Frege. Mais cette analyse semble animée par un style philosophique bien classique dont on trouve l'équivalent chez Aristote. Si « on résout les arguments qui sont de véritables raisonnements en les détruisant », on résout « ceux qui sont seulement apparents en faisant des distinctions » (*Réfutations sophistiques,* 1766).

Mais que doit signifier une analyse des significations, après la découverte des antinomies logiques, plus sérieuses qu'apparentes, et qui, de plus, semblent mettre en doute le sens commun, voire la logique elle-même qui, disait Poincaré, « n'est plus stérile [puisqu'] elle engendre des contradictions » ? La théorie des types rémédiait à cette fâcheuse situation dans la mesure où elle permettait de faire disparaître ce qui semblait à Russell la racine des contradictions : les définitions imprédicatives qui distinguent les éléments d'un ensemble par des termes qui se réfèrent à cet ensemble, autrement dit, qui emploient le quantificateur universel sans précaution. Sans entrer dans des détails de peu de sens pour ce propos, on dira, étant donné une fonction propositionnelle $\varphi(x)$, qu'il existe un certain nombre de valeurs de x pour lesquelles cette *fonction a un sens,* c'est-à-dire est vraie ou fausse. Soit a l'une de ces valeurs, alors $\varphi(a)$ est une proposition vraie ou fausse. Mais $\varphi(x)$ peut également être toujours vraie ou toujours fausse. Ainsi, « si x est humain, x est mortel » est toujours vraie, mais « x est humain » l'est seulement quelquefois. Donc une *fonction propositionnelle* (on désigne par là une *formule ouverte,* c'est-à-dire qui contient des variables libres ; *cf.* le chapitre *Logique*) peut être considérée de trois manières : soit en substituant une constante à une variable, soit en affirmant toutes les valeurs de la fonction, soit en affirmant certaines valeurs seulement. Pour éliminer les contradictions dans le cadre du « sens commun », il faut, lorsque j'affirme toutes les valeurs d'une fonction $\varphi(x)$, que le domaine des valeurs de x soit bien déterminé, « c'est-à-dire, écrit Russell, qu'il doit y avoir une totalité quelconque de valeurs possibles de x. Si maintenant je crée, ajoute-t-il, des valeurs définies en termes de cette totalité, la totalité paraît, de ce fait, agrandie, et c'est pourquoi les nouvelles valeurs qui s'y réfèrent se référeront à une totalité agrandie. Mais puisqu'elles doivent être comprises dans cette totalité, celle-ci ne peut jamais les rattraper. C'est comme d'essayer de sauter sur l'ombre de votre tête » (*Histoire de mes idées philosophiques,* Gallimard, 1961, p. 102). Lorsque le menteur dit : « Je mens », s'il dit par là : « Tout ce que j'affirme est faux », cette assertion se réfère alors à la totalité de ses assertions, et c'est seulement dans cette mesure que se produit le paradoxe. Or, il est possible de définir une hiérarchie de propositions telle que celles du premier ordre ne se réfèrent pas à la totalité des propositions du premier ordre. Une fonction propositionnelle ne peut plus, ainsi, se prendre elle-même comme valeur de son argument. Le menteur pourra alors bien dire : « J'affirme une fausse proposition de premier ordre qui est fausse. » Mais cette proposition, dit Russell, « est elle-même une proposition de second ordre. Il n'affirme donc ainsi aucune proposition de premier ordre. Ce qu'il dit est ainsi tout simplement faux, et l'argument selon lequel c'est également vrai tombe ». Ainsi, la théorie des types s'attaque à la raison des paradoxes : elle élimine leur risque en supprimant la possibilité de cette « référence à soi réflexive » qui comprend « comme membre d'une totalité, quelque chose qui se réfère à cette totalité qui ne peut avoir un sens défini que si la totalité est déjà fixée ». L'élimination des paradoxes en résulte, car, désormais, il est impossible de confondre la relation d'appartenance d'un élément à un ensemble et la relation d'inclusion d'une partie d'un ensemble dans un autre. Les quantificateurs ne pourront se rapporter qu'à un type déterminé de variables, jamais à tous les types possibles.

Toutefois, l'ensemble des mathématiques n'est pas encore sauvé. Pour ce, il a fallu introduire le fameux axiome de réductibilité dont Wittgenstein disait que, s'il était vrai, ce serait par hasard qu'il le serait. Cet axiome, permettant dans la version ramifiée des types (nous n'examinons pas toutes les versions de cette théorie) de considérer toute fonction propositionnelle d'un niveau quelconque comme ayant la même extension qu'une fonction propositionnelle élémentaire, semble réintroduire par une porte de derrière des définitions imprédicatives.

Quoi qu'il en soit, la contribution du logicisme à l'analyse du langage est décisive. La solution aux antinomies a été à l'origine également de la théorie des descriptions, dont le point central est qu'une expression « peut contribuer au sens d'une phrase sans posséder aucun sens quand elle est seule ».

Par là aussi, elle satisfaisait un requisit de l'analyse frégéenne : qu'on doit chercher ce que les mots veulent dire non pas isolément mais pris dans leur *contexte*. Ce principe du contexte, produit par et pour une réflexion philosophique, n'a pas été seulement déterminant dans la philosophie analytique, qui a pris ses distances envers la logique, mais reste parfois actif dans une réflexion purement technique, qui s'est libérée de l'horizon logiciste, sans le renier. Car les théories des types essayées par H. Wang et Lorenzen ont rendu caduques les vieilles objections, et l'idée d'une distinction entre les différentes couches du langage a donné à une vieille idée philosophique des interprétations exactes intéressantes, au-delà des frontières logicistes. En effet, l'idée, par exemple, de distinguer entre catégories différentes d'objets n'est pas absente même de la fondation de la théorie des ensembles de Newman-Bernays où deux univers disjoints sont admis, celui des classes et celui des ensembles. Mais il y a davantage, il est possible aujourd'hui de passer d'une théorie pluricatégoriale (qui admet plus d'une sorte d'objets fondamentaux) à une théorie monocatégoriale (qui n'admet qu'une seule sorte d'objets). Ce qui est une manière, non seulement d'apprécier la valeur interne d'un style d'analyse logique, mais aussi d'être plus sûr quant à sa validité, à un moment où les philosophies se sont si fondues dans les techniques qu'il faut une multiphilosophie pour comprendre ces techniques, enracinées, pour les plus récentes, dans une pensée logique encore grosse d'avenir.

Le formalisme

Mais il fallait que le logicisme se rendît compte qu'« avec l'infini commence la véritable mathématique », autrement dit que l'essentiel était, après les antinomies, de fonder l'analyse telle que la pratique le mathématicien, et en sorte que son travail pût se poursuivre dans la sérénité. « Dans sa fondation de l'analyse, Weierstrass, disait **Hilbert,** a accepté sans réserve et utilisé à maintes reprises ces formes de déduction logique dans lesquelles le concept de l'infini entre en jeu, comme lorsqu'on traite de *tous* les nombres réels ayant une certaine propriété ou lorsqu'on démontre qu'il *existe des* nombres réels avec une certaine propriété. » Ce que Hilbert dit de Weierstrass, il pouvait le dire également de Dedekind, de Cantor, et d'autres mathématiciens qui ont vu surgir, sous leurs yeux, la nécessité de considérer des totalités infinies, autrement dit de manipuler l'*infini actuel*.

Légitimer donc l'infini actuel, telle doit être la tâche immédiate du mathématicien. Cette tâche s'impose comme un choix à la pensée mathématique. Car elle ne semble pas la seule offerte. Outre qu'on peut bannir purement et simplement l'infini actuel, on peut aussi l'admettre de diverses manières. Ce sont donc plusieurs voies qui s'offrent.

La première en date est celle de **Zermelo.** Elle puise à deux sources :

— Zermelo s'inspire d'abord de Dedekind, qui avait élaboré une théorie des systèmes comme soubassement logique de sa théorie des nombres. Mais ce n'est pas seulement le matériau qui inspire Zermelo. Une idée plus profonde de Dedekind, plus profonde que le fait que les entiers naturels ou finis sont fondés sur l'existence d'ensembles infinis dénombrables, est qu'il utilise des *axiomes* (généralement attribués à Peano) pour définir ces entiers et les considère comme des conditions de structure sur des objets de nature non spécifiée. Ce type de méthode, il l'a lui même introduit dans sa *théorie des idéaux* (on appelle « idéal » une sous-structure de la structure d'« anneau », qui permet notamment le passage à la structure « d'anneau quotient », de même qu'un sous-groupe d'un groupe [commutatif] permet de considérer le « groupe quotient »), qui l'avait déjà opposé à Kronecker, dans les débuts de la seconde moitié du XIXe siècle.

— Zermelo a pris, ensuite, une leçon décisive dans les *Fondements de la géométrie* de Hilbert qui est une réalisation plus explicite encore de l'idée d'axiomatique moderne. Au lieu de définir l'objet dont traite une discipline mathématique, on part d'un certain nombre d'axiomes écrits à l'aide de symboles donnés, par exemple celui de la relation d'appartenance par le truchement de laquelle est signifiée l'existence de certains ensembles. L'ensemble des axiomes définit ainsi implicitement le concept d'ensemble, comme dans l'axiomatique formelle de la géométrie où Hilbert n'admet pas d'avance les relations primitives comme déterminées dans leur contenu, car elles ne le sont que par les axiomes, et ne sont utilisées que par ce qui en est formulé expressément par les axiomes. D'une certaine manière, les axiomes ont ainsi force de définition, en ce sens que, au lieu de définir les notions de « point » ou de « droite » et de « plan » comme le faisait Euclide, non sans embarras, ils définissent des lettres appartenant à trois domaines d'individus distincts d'un calcul *logique* (dans le cas où l'on pousse la formalisation plus loin). Il en est à peu près de même de la théorie des ensembles de Zermelo, qui ne laisse plus de place aux antinomies.

Dans l'horizon de ce nouveau style, il n'est plus directement question de fonder les axiomes, dont celui de l'infini. Ce qui importe aux mathématiciens-techniciens, la réduction de la théorie des ensembles à quelques principes précisément formulés et soigneusement distincts de toute préoccupation étrangère, est par là assuré. En particulier, la fondation de la théorie axiomatique des ensembles semble être du ressort d'un domaine différent : celui de la métamathématique.

Mais cette solution technique implique deux attitudes philosophiques.

L'une, implicite, a occupé aussi peu les philosophes que les mathématiciens, à l'exception du seul Desanti : c'est le problème des conditions d'une « reproduction » de théorie, qui est précisément une reproduction d'axiomes s'effectuant selon certaines modalités dialectiques du passage de l'implicite à l'explicite (*Idéalités mathématiques,* Éd. Seuil, 1968).

L'autre attitude concerne la philosophie explicite qui s'est immédiatement greffée sur le problème des théories axiomatiques et de leur fondation : le formalisme, précisément.

En gros, le formalisme est d'abord une conception des mathématiques qui donne toute son importance à la notion de *symbole,* à laquelle elle tend à réduire la vieille notion d'être mathématique, qui n'y est plus de mise. « La condition préalable à l'application des raisonnements logiques... c'est que quelque chose, écrivait Hilbert, soit donné à la représentation : à savoir certains objets concrets, extralogiques, qui sont présents dans l'intuition en tant que données vécues, immédiates, préalablement à toute activité de pensée. Pour que le raisonnement logique soit doué de solidité, il faut que l'on puisse embrasser ces objets du regard dans toutes leurs parties, et que l'on puisse reconnaître par intuition immédiate, en même temps que ces objets eux-mêmes, comme des données qui ne se laissent pas réduire à quelque chose d'autre, ou qui, en tout cas, n'ont pas besoin d'une telle réduction, la façon dont ils se présentent, dont ils se distinguent les uns des autres, la façon dont ils se suivent ou dont ils sont rangés les uns à côté des autres. Telle est la position philosophique fondamentale que je considère comme essentielle pour les mathématiques aussi bien que pour toute espèce de pensée, de compréhension et de communication scientifique. *En mathématiques, en particulier, l'objet de notre examen, ce sont les signes concrets eux-mêmes dont la forme nous apparaît immédiatement avec évidence,* conformément à notre position fondamentale, et demeure parfaitement reconnaissable. » Autrement dit, comme l'écrira encore Tarski, on peut considérer « les énoncés comme des inscriptions et par suite comme des corps physiques concrets »

(textes cités par R. Martin, *Logique contemporaine et Formalisation,* P. U. F., 1964, pp. 28-29). Au commencement, disait encore Hilbert, il y a le signe. Comme si Hilbert supposait nécessaire, tout simplement, de savoir lire et écrire, inscrire et discerner des signes, pour pratiquer une science.

Le formalisme s'est, ensuite, assigné la tâche de formaliser toutes les mathématiques, tâche qu'a rendue possible leur axiomatisation.

Toutes les mathématiques, au moins jusqu'au niveau atteint par les *Principia mathematica* de Russell et Whitehead, doivent pouvoir être reconstruites en *systèmes formels,* à partir de certains axiomes et à l'aide d'un certain nombre de règles d'inférence ; cela consiste donc à monter pour ainsi dire des mécanismes logiques où n'intervient jamais le sens des formules sur lesquelles on opère.

Hilbert va plus loin. La notion de système formel, établie ainsi indépendamment de toute représentation, ne suffit pas. Il faut s'assurer que l'application des règles d'inférence aux axiomes ne conduit jamais à une contradiction, c'est-à-dire qu'elle ne permet jamais de dériver à la fois $a = b$ et $a \neq b$, ou $1 = 2$. Cette assurance doit faire l'objet d'un métathéorème d'impossibilité soustrait au moindre doute ; c'est un élément d'une métathéorie qui étudie les démonstrations formelles et qui est donc une *métamathématique* ou *théorie de la démonstration.* Dans cette dernière théorie, Hilbert emploie une arithmétique *finitiste,* mettant en œuvre « un raisonnement direct, pourvu de contenu, qui s'accomplit en expériences de pensée sur des objets donnés à l'intuition en dehors de tout présupposé axiomatique ». Il s'agit en fait d'une arithmétique concrète immédiate, acceptable par tous les mathématiciens, y compris ceux qui refusaient résolument de considérer une collection infinie comme une donnée actuelle. C'est donc une arithmétique minimale et philosophiquement neutre, mais qui n'a été élaborée que sur les principes philosophiques d'une conception formaliste *muée en technique de formalisation,* d'autant plus fondée qu'il y a un « parallélisme absolu, comme disait J. Herbrand (*Écrits logiques,* P. U. F., 1968, p. 37), entre le raisonnement mathématique et les combinaisons de signes », parallélisme qui rend naturelle l'étude des systèmes de signes pour eux-mêmes. Mais ce formalisme ne promet pas seulement de concilier toutes les écoles nées du refus de l'infini actuel, de son adoption, ou d'attitudes nuancées intermédiaires, il a l'ambition de restaurer l'idée de rigueur absolue historiquement attachée aux mathématiques. Il prétend fonder, en dernière analyse la mathématique existante par des moyens si sévèrement sélectionnés que leur refus est inadmissible. C'est pourquoi Herbrand écrivait encore qu'il cherche seulement à examiner des théories existantes et étudie les caractères des propositions qui y sont vraies ; il ne prend pas part aux discussions que celles-ci soulèvent ; il ne cherche pas à les départager. Bref, le formalisme apparaît philosophiquement neutre, ce qu'il est incontestablement dans ses intentions. Faut-il penser que les théorèmes de Gödel (1931), dont résulte l'impossibilité de réaliser le programme de Hilbert, montrent en même temps l'impossibilité du neutralisme philosophique impliqué par le finitisme conçu comme moyen de sauvegarder, en dernier ressort, l'ensemble de la mathématique ? De simples restrictions finitistes ne semblent pas en mesure de démontrer avec précision la légitimité du point de vue formaliste, qui s'appuie, en fin de compte, sur la croyance qu'aux énoncés intuitifs correspondent toujours des énoncés dérivables des axiomes d'un système formel. Ainsi était détruite, à sa base, une idée chère à Hilbert mais également partagée, quoique sous une autre forme, par les logicistes. Ce qui est rendu impossible par là, c'est plus qu'une philosophie, c'est l'homogénéité philosophique des mathématiques.

L'intuitionnisme

Gödel est parvenu à ses résultats en se servant, dans son travail, de *tout* ce qu'alors on pouvait savoir. Il a abattu les cloisons qui séparaient artificiellement les théories mathématiques différentes, en appliquant une mathématique intuitive à l'analyse de la mathématique formelle, en élargissant l'idée du finitisme, de celle d'arithmétique *minimale* susceptible de battre les intuitionnistes sur leur propre terrain à l'aide d'un armement encore plus limité que le leur, à celle d'une arithmétique intuitive, c'est-à-dire à la considération du programme intuitionniste. Avec Gödel, la réflexion formaliste et la réflexion intuitionniste, comme les résultats obtenus par l'école logiciste, perdent leur statut proprement philosophique pour n'être considérées qu'en qualité d'*outils.* Si cette position épistémologique, implicite dans les résultats de Gödel, n'a pas été extrêmement favorable au programme de Hilbert, elle a permis de consacrer l'importance de l'intuitionnisme au point de vue classique même, dès le moment que ce point de vue quittait le pragmatisme mathématicien pour s'intéresser aux problèmes des fondements des mathématiques.

Qu'est-ce alors que l'intuitionnisme ? C'est une conception des mathématiques développée par **Brouwer** et son école (dont les plus illustres sont Beth et Heyting). D'accord avec le logicisme, il reproche aux axiomaticiens de construire des édifices de rhétorique à la signification non assurée, mais contre les logicistes il tend à établir que les mathématiques constituent une activité spécifique et indépendante de l'esprit, irréductible, en tant que telle, à la logique pure. Cette attitude originale aboutit, derrière le refus de la théorie des ensembles de Cantor et de Zermelo, à ne rien admettre qui mette en œuvre l'idée de totalité infinie actuelle. Anticipant l'idée que l'évidence du fini ne permet pas de procurer à l'infini une évidence qu'il n'a pas, de par sa nature, elle donne une nouvelle portée à l'ancien intuitionnisme ; elle donne raison à Kronecker contre Cantor et Dedekind et rejette ainsi les principaux résultats admis au fondement de l'analyse, tels que le théorème de Bolzano-Weierstrass. Bref, il y a là de quoi ouvrir la voie à une nouvelle réflexion sur les notions d'*objet,* d'*existence* et de *preuves mathématiques.*

On a eu trop tendance à considérer qu'un objet mathématique était donné dès le moment où son caractère non contradictoire était établi. Pour l'intuitionniste, le critère purement logique de non-contradiction ne suffit pas. De manière plus générale, il ne suffit pas de la compatibilité, dans le sens de la non-contradiction à l'intérieur d'une théorie mathématique, pour établir l'existence de ses objets. La vérité n'a pas seulement à être non contradictoire. En mathématiques, il faut plus ; il faut une *construction,* et c'est la possibilité de celle-ci qui est décisive. Cela rappelle, certes, une attitude plus traditionnelle : Descartes et Kant étaient convaincus de l'inadéquation de la logique formelle aux objets mathématiques, dans la mesure où, pour Descartes, elle ne suffit pas à les inventer, et dans la mesure où, pour Kant, la mathématique est en elle-même une activité synthétique pure, irréductible à la logique, qui est par essence analytique. Autrement dit, le raisonnement mathématique n'est pas seulement l'enchaînement d'inférences formelles : il procède, comme disait Kant, « dans l'intuition de l'objet ». Mais le néo-intuitionnisme va plus loin : existence signifie *constructibilité.* Ce qui n'est pas constructible est donc rejeté, parce que non intuitif. C'est pourquoi l'axiome de choix est rejeté par les intuitionnistes de différentes obédiences, et cela indépendamment de la question de savoir si cet axiome est compatible avec les autres. La compatibilité en soi, ou d'une théorie, garantirait aussi peu l'existence que, au point de vue intuitionniste, le manque de preuve, l'innocence d'un criminel. C'est donc positivement que l'objet mathématique est caractérisé, et intrinsèquement qu'il est intuitif. Mais l'idée de construction domine toute la production intuitionniste ; elle vaut également pour le concept de déduction : une formule n'est *prouvée* que si elle résulte d'une construction.

On voit par là que l'intuitionnisme n'est pas hostile à la logique, mais il lui assigne le moment précis de son intervention : quand il s'agit d'analyser le langage d'une théorie mathématique déjà avérée. Cette logique peut être formalisée dans le cadre intuitionniste et l'a été effectivement par Heyting.

Par exemple, l'implication $A \rightarrow B$ y peut être prouvée comme une construction transformant toute preuve de A en une preuve de B ; ou bien, selon Kolmogorov, comme le problème qui consiste à réduire le problème de B à celui de A, ce qui s'accorde en soi avec la conception intuitionniste de l'identité entre un énoncé mathématique et sa construction. De même, il y a une idée intuitionniste de l'arithmétique des entiers. Tout entier est confondu avec sa construction, et la suite n'est supposée donnée



Lauros - Giraudon

▲ ***Première page avec frontispice tirée d'un ouvrage d'Histoire naturelle, rédigé en Italie au XVe siècle.***

qu'en ce sens, et non comme totalité infinie actuelle. La suite des entiers joue, aux yeux de Brouwer, un rôle fondamental ; elle est à la racine des êtres mathématiques ; le procédé qui permet de la parcourir dérive d'une intuition primitive où intervient l'écoulement du temps. Enfin, avec les entiers intuitifs à la base, on peut définir des types de suites, dont les suites de choix. Celles-ci permettent de construire un continu qui n'est pas, comme celui de l'analyse arithmétisée, composé de points qui sont les éléments d'un continu linéaire ; il embrasse seulement des parties, divisibles sans limite, sans cesser d'être continues. C'est un « milieu de libre devenir », sur la base duquel il est aussi possible d'élaborer une théorie des nombres réels ou une topologie.

Cependant, s'il y a un aspect par lequel l'intuitionnisme est devenu célèbre, c'est la mise en doute des principes logiques tels qu'ils étaient enseignés par une tradition qui remonte à Aristote. L'intuitionnisme rejette, comme on le sait, la validité dans l'infini du *tertium non datur* ou tiers exclu (principe suivant lequel une proposition est soit vraie, soit fausse). Une démonstration appuyée sur ce principe n'emporterait pas l'adhésion. L'alternative « deux points doivent être distincts ou confondus » n'est pas contraignante, car il existe un couple de points qui ne sont ni l'un ni l'autre, comme l'illustre le développement décimal de π où l'on désigne par R le rang décimal où commence pour la première fois la suite 1 2 3 4 5 6 7 8 9. Si nous définissons un nombre réel tel que $\rho = \pi + 10^{-k}$, on peut écrire de proche en proche le développement de ρ. En particulier, $\rho = \pi$ est vrai si une démonstration générale établit que, dans le développement décimal de π, la suite précitée n'apparaît jamais. Bien que ρ soit univoquement déterminé, des principes constructifs ne permettent pas d'affirmer s'il est ou bien différent de π, ou bien égal à π : s'il est différent, c'est qu'on a trouvé le nombre *k*, et s'il est égal, c'est que l'on dispose de la démonstration générale mentionnée. Le tiers exclu ne s'applique pas. De plus, selon Brouwer, l'application sans limite du tiers exclu suppose la résolubilité de tout problème mathématique. Le programme de Hilbert tient implicitement compte de cette critique dans l'idée d'une arithmétique minimale. C'est déjà là un lien entre théorie de la démonstration et intuitionnisme. En effet, pour les démonstrations de non-contradiction, on pare au défaut de puissance des méthodes strictement finitistes en atténuant le point de vue finitiste par l'adjonction de méthodes démonstratives nouvelles. Par exemple, on se fie à celles des intuitionnistes pour disposer d'une évidence maximale, faute d'une arithmétique minimale ; c'est la voie indiquée par Herbrand et dans laquelle s'engage Gentzen dès 1936 pour démontrer la non-contradiction d'une importante partie de l'arithmétique élémentaire (en utilisant le principe d'induction transfinie jusqu'à ε_0). La relation étroite entre l'arithmétique classique et l'intuitionnisme s'énonce précisément dans le théorème de Kolmogorov-Gödel, qui nous apprend que la non-contradiction de l'arithmétique classique est (logiquement) *équivalente* à celle de l'arithmétique intuitionniste. Dès lors qu'il existe ainsi une *interprétation* intuitionniste d'*une* théorie classique, celle de l'arithmétique, la voie est ouverte à une extension possible (?) de ce fait à d'*autres* théories, et la nécessité posée d'une réflexion à poursuivre sur la notion de démonstration.

Les différentes attitudes à l'égard du problème des fondements des mathématiques n'ont donc plus aujourd'hui leur signification d'origine. Si entre logicistes et intuitionnistes s'est élevée une querelle de priorité, le débat entre formalistes et intuitionnistes s'enracine davantage dans le problème des fondements proprement dits. L'intuitionnisme a évolué et n'a pas été aussi peu fécond qu'on le craignait, tandis que le formalisme est revenu sur des promesses trop amples, tout en ayant dû étendre ses moyens, pour la réalisation du programme de Hilbert révisé, au-delà de ce qui était intuitivement assuré pour l'intuitionnisme strict. Tout s'est donc passé comme si les diverses attitudes n'avaient fait que s'offrir les unes aux autres des outils indispensables à la réalisation (plus ou moins limitée) de leurs exigences.

Philosophie et épistémologie des mathématiques

Les conceptions les plus philosophiques des mathématiques se sont muées en techniques si précises que la réflexion philosophique y est devenue sinon impossible, du moins trop exigeante pour qu'il soit pratiquement possible de concilier la pertinence et l'intelligibilité philosophique. Car c'est le philosophe qui est aujourd'hui un profane. Dans la mesure où les mathématiques ne s'offrent plus comme discours à dépasser vers un discours philosophique ultime et fondateur, on peut se demander si quelque philosophie peut encore avoir sa place *au voisinage des* mathématiques.

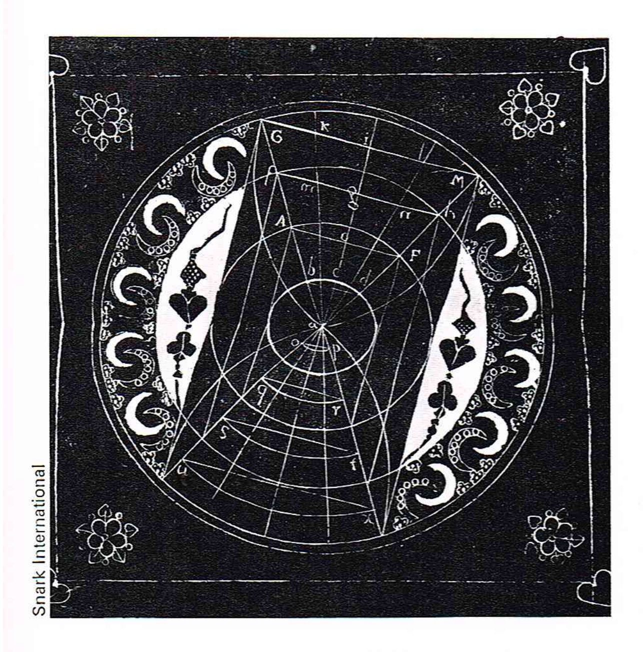
Problèmes philosophiques des mathématiques actuelles

La première tâche qui s'offre à une telle interrogation, c'est de préciser de quelles mathématiques il s'agit. La totalité intégrale des mathématiques, construite de proche en proche, capable d'éclairer le pur et l'appliqué, d'expliquer la nature ultime des êtres et des démarches, c'est sans doute un beau rêve, mais il ne semble pas avoir grand sens. Qu'en est-il alors de l'*unité* des mathématiques ?

Longtemps, les mathématiques ont vécu sur une distinction bien nette entre le discret et le continu, séparés par un fossé qu'*on n'a comblé jusqu'ici que pour le creuser davantage*. Depuis Descartes, on n'a cessé de jeter les ponts entre les deux domaines. D'abord un parallélisme de structure entre la géométrie et l'algèbre, qui s'est de plus en plus précisé. Ensuite, le développement de l'analyse, c'est-à-dire le « saut » dans des considérations où le concept d'infini joue un rôle. Enfin, l'arithmétisation de l'analyse, c'est-à-dire l'élaboration purement arithmétique du contenu linéaire, qui a permis à Hilbert de réduire la géométrie elle-même, ou plutôt les types de géométrie à l'arithmétique des réels. Le tour de force admiré consacre l'unité de la grandeur et du nombre, indique que ce qui s'applique au calcul des nombres réels s'applique également au calcul avec des segments de droite. Or, c'est la méthode axiomatique qui a procuré ces résultats.

Cette méthode est féconde par elle-même, mais aussi par les problèmes qu'elle pose. Elle conduit aux idées de formalisation et de système formel. Mais elle permet aussi d'élucider et d'expliciter la nature de types abstraits de structure, de poser les problèmes au plus haut degré de généralité et de précision, de montrer l'intérêt de l'appareil du langage et de la logique, la nécessité de distinguer le système et ses interprétations, puisque l'étude d'un système formel induit celle d'une métathéorie qui, si elle est précise, doit avoir les moyens de nommer sans confusion les termes et les formules du système qu'elle étudie.

Mais cette unité méthodologiquement assurée par l'axiomatisation n'est pas aussi entière au niveau le plus profond, c'est-à-dire de la fondation justifiée du continu, où intervient l'infini, sur le discontinu qui peut être infini potentiel, mais qu'il n'y a aucune nécessité à considérer


Snark International

◀ *Dessin original tiré de* **Assiculi... adversus mathematicos atque philosophos** *de* **Giordano Bruno** *(1548-1600), penseur et écrivain du XVIe siècle. Il réussit à propager et à défendre le* **Système Copernic,** *mais sa cosmologie lui valut de périr sur le bûcher en 1600.*

comme ensemble achevé : ne suffit-il pas de dire, comme Poincaré, que « quand je parle de tous les nombres entiers, je veux dire : tous les nombres entiers qu'on a inventés, et tous ceux que l'on pourra inventer un jour... et c'est ce « que l'on pourra » qui est l'infini »? On a vu qu'il n'est pas possible de joindre l'infini à partir de finitismes stricts. Suffit-il d'adjoindre aux moyens intuitivement assurés, qui ne reposent pas sur l'infini actuel, des énoncés axiomatiques tels que ceux de l'infini ou du choix par simple droit et autorité de législation axiomatique? Est-ce une raison suffisante pour justifier cela que d'alléguer, ce qui est le recours constant du « working mathematician », qu'il nous faut produire des fondations susceptibles de reproduire l'acquis classique, dont personne ne met en doute la signification pour la mathématique aussi bien appliquée que pure? Est-ce là le signe du caractère arbitraire des choix classiques, qui ne justifieraient alors que des développements liés à leur tour aux développements et aux exigences de domaines étrangers à la mathématique? Non, sans doute. Car il semble précisément que les parties les moins liées aux domaines étrangers sont celles qui ont un caractère « spéculatif », et qui sont le fruit de l'imagination logique. Cela empêche pour le moins tout jugement précipité, toute philosophie arrêtée des mathématiques qui mobilisent chaque fois à nouveau une réflexion philosophique. Car une chose est certaine : c'est que les mathématiques sont en perpétuelle interrogation sur leur organisation, les rapports entre leurs différentes parties, leurs différents résultats, les meilleures manières de les agencer. Elles sont toujours en réorganisation. Mais elles ne cessent pas de se poser des problèmes au sujet de leur validité. C'est pourquoi les théories de la démonstration ne tendent pas à donner une réponse définitive à un problème définitivement formulable, mais suivent le mouvement de cette liberté créatrice, même si sa création n'est point l'écoulement continu et sans heurts des propositions et des théorèmes. Les mathématiques alors ne prouvent pas, pour reprendre ici une expression bien connue, le mouvement en marchant, mais elles produisent les principes de leur propre mouvement. Qu'est-ce à dire sinon qu'il est imprudent, non seulement de philosopher sur les mathématiques à partir d'une philosophie atrophiée, mais aussi de croire qu'il n'y a pas de rapport entre philosophie mathématique et mathématique, puisque cette croyance fut celle du programme de Hilbert, lequel a été, précisément, c'est-à-dire mathématiquement, réfuté?

Épistémologie et histoire des mathématiques

S'il existe donc une réflexion philosophique, techniquement armée, capable de maintenir dans son interrogation l'exigence de sonder la validité mathématique dans une théorie de la preuve, inachevée et prometteuse, susceptible de réaliser des progrès suffisamment significatifs pour avoir des répercussions sur le cadre logique usuel, dont elle révèle l'extrême simplicité par rapport aux constructions multiples qu'on peut y réaliser, cette réflexion se maintient et se cultive au sein de cette mathématique supérieure qui s'occupe (encore et toujours) [*cf.*, par exemple, maints articles de G. Kreysel] des fondements. Le philosophe, c'est le mathématicien que la mathématique suscite pour débloquer une problématique, dépasser une situation figée par des attitudes d'école, dégeler un point de vue ancré dans la technique. Le philosophe, c'est le mathématicien soucieux de la signification des démarches qu'il effectue, qu'il engage sur la base d'une analyse du matériau auquel il donne une nouvelle perspective, ou qu'il gagne à une nouvelle fonction. Un travail de pensée gît souvent sous l'écorce conceptuelle. Et c'est du travail de philosophe, au sens propre, que sa mise en chantier en vue de sa restitution, non pas en tant qu'activité mentale inaccessible, mais en tant que téléologie intellectuelle dont on doit expliciter la démarche et les moyens dans une configuration conceptuelle donnée. Parfois ce travail est déjà indiqué par le mathématicien : posant qu'il existe deux théories des ensembles, l'une qui admet l'axiome de constructibilité, l'autre qui le rejette. Quand on admet la constructibilité, on admet, évidemment, un certain nombre de conséquences et de propriétés qui en résultent, dont le fait que l'axiome est compatible avec les autres axiomes de la théorie. Cette consistance relative en établit la vérité dans le domaine des ensembles constructibles. D'où la nécessité d'une analyse de la notion de constructibilité. Mais cette analyse n'est pas conduite conceptuellement. Le concept se déploie dans la technique. Car c'est un travail nécessairement technique, lui-même impossible sans analyse conceptuelle, sans élucidation de la notion de constructibilité dans ce contexte. C'est-à-dire que les techniques mathématiques ne se réduisent plus toutes à des problèmes qu'on peut confier à des sous-traitants. Elles fournissent les conditions minimales de la rationalité mathématique.

C'est précisément ce qui justifie l'analyse épistémologique de moments mathématiques privilégiés, qu'on poursuit en France, depuis les analyses par trop philosophiques d'A. Comte jusqu'aux exigences plus raffinées de Bachelard, de Cavaillès et de Desanti. Leur justification est exprimée par L. Brunschvicg, qui pensait que c'est le développement de la mathématique « qui nous instruit des fonctions de l'intelligence », ce qui demeure vrai si l'on prend soin de briser les cadres factices des facultés, de briser même ce sur quoi L. Brunschvicg bâtissait : le cadre factice de « l'intelligence occidentale », de l'idée ambiguë de développement, de l'illusion d'une critique philosophique extérieure. Mais l'idée d'un chemin qui n'est ni l'étude de l'activité mentale mathématicienne, ni la réflexion outillée sur l'idée de construction, de preuve, voire de déduction, et qui permet une caractérisation intrinsèque du « réel » mathématique, ne peut être interdite.

A une condition cependant, qui doit être effectivement remplie : que l'épistémologue participe à la fois « du mouvement de l'intelligence et de la rigueur logique ». Qu'il y participe : c'est-à-dire qu'il s'interdise de fixer des notions mobiles ou de mobiliser des notions inertes. Qu'il saisisse que la logique mathématique elle-même est un des mouvements de l'intelligence.

Ce mouvement de l' « intelligence » a été caractérisé par Cavaillès comme un mouvement « dialectique ». Cela s'entend d'abord négativement : car ni l'expérience au sens physique, ni aucun *a priori* logique ne saurait fonder les mathématiques. Ni non plus, à vrai dire, aucune genèse empirique, car « l'enfant devant son boulier est mathématicien, et tout ce qu'il y peut faire est mathématique». La mathématique est dès son «commencement» mathématique à part entière. Mais son devenir effectif reflète une structuration où l'on distingue des aspects paradigmatiques et des constructions thématiques. Une construction paradigmatique est, par exemple, le mouvement formel, « le moment de la variable » où, « en remplaçant les déterminations d'actes par la place vide pour une substitution, on s'élève progressivement à un degré d'abstraction qui donne l'illusion d'un formel irréductible ». Tel serait le moment hilbertien pris dans son mode de réflexion sur les systèmes formels et leurs interprétations. Toutefois, « la formalisation n'est réalisée que lorsque au dessin des structures se superposent, systématisées, les règles les régissant. La *thématisation*

prend son départ dans l'enchaînement saisi cette fois dans son vol, trajectoire qui se mue en sens » (*Sur la logique et la théorie de la science,* pp. 29-30). Style second qui est le moment métamathématique où le thème principal est la fondation du mouvement démonstratif lui-même.

C'est dans cette voie que s'est engagé J. T. Desanti, convaincu qu'une épistémologie des mathématiques doit rompre toute attache avec les discours philosophiques en « s'installant dans le contenu des énoncés scientifiques. Pour les mathématiques comme pour les autres sciences, il faut « ou se taire ... ou bien en parler de l'intérieur, c'est-à-dire en les pratiquant » (*Philosophie silencieuse,* Seuil, 1969, p. 108). Mais il y a des problèmes épistémologiques, c'est-à-dire des systèmes d'expressions appartenant à un domaine théorique donné et qui exigent, à un moment du développement de cette théorie, une attention particulière, la mise en chantier d'une réflexion indispensable, soit pour lever un obstacle à l'intelligibilité de la théorie, soit pour parvenir à une intelligibilité supérieure. Ainsi, la théorie naïve des ensembles a-t-elle entre 1900 et 1908 posé un certain nombre de problèmes :

— exigence d'une démonstration de la proposition « Tout ensemble peut être bien ordonné »;

— exigence de restrictions sur le concept d'ensemble pour éviter les paradoxes;

— exigence enfin de déterminer le statut de l'axiome de choix, indispensable à la démonstration du théorème de bon ordre.

Or, si la première et la troisième exigences ont un caractère régional, la seconde implique le programme d'une axiomatisation au sens abstrait qui doit être la reproduction de l'édifice ensembliste tout entier muni de garanties anti-contradictoires, c'est-à-dire, en fait, l'élaboration d'un langage rigoureusement construit. Nous avons déjà remarqué plus haut que Desanti s'est attaché à l'analyse du mouvement complexe d'axiomatisation dans les *Idéalités mathématiques.* Mais on peut envisager une épistémologie qui, au-delà des raffinements intrinsèques, s'intéresserait aux variétés stylistiques de la construction mathématique, qui tâcherait d'intégrer l'individuel, le processus concret du travail dans l'« œuvre » mathématicienne.

Ainsi, comme le remarque G. G. Granger (*Essai d'une philosophie du style,* A. Colin, 1968, p. 20), on peut introduire la notion de nombre complexe de plusieurs manières sans modifier les propriétés opératoires caractérisant la structure algébrique du « corps des complexes ». La représentation trigonométrique fait intervenir un angle, l'argument, et un nombre réel, le module. Le mathématicien danois du XVIIIe siècle Wessel l'a proposé dans l'intention d'établir un calcul qui porte à la fois sur les grandeurs et les « directions ». L'être mathématique ainsi constitué peut être envisagé de deux manières : soit comme élément statique, ou vecteur, soit comme un opérateur appliqué à des vecteurs. Le premier cas suggère un passage naturel des coordonnées polaires aux coordonnées cartésiennes. Mais le second suggère, grâce à l'apport intuitif de l'image géométrique, une construction immédiate des lois de la multiplication des nombres complexes. Toutefois, d'autres concepts sont possibles du « corps » des complexes : soit comme matrice carrée régulière d'un certain type, soit plus abstraitement comme corps d'extension des réels. Ces différentes façons constituent des faits de style. Ces faits peuvent être à l'origine de véritables variations conceptuelles et peuvent déterminer l'orientation d'un concept. Le « style » révèle donc un aspect du travail mathématique, « une dialectique de son développement interne » qui vient enrichir l'épistémologie de nouvelles nuances, de nouveaux aperçus sur ce travail.

Mais ce type d'épistémologie, aussi proche qu'il soit du travail mathématique, comme le souhaite Desanti, reçoit la mathématique constituée. Ce savoir accompli, le philosophe n'a donc pas à l'énoncer. Mais seulement à détruire le discours aveugle et second qui l'accompagne, à déchiffrer, au plus près de l'activité mathématique, ce qui enchaîne et éclaire les motivations qui lui sont propres. Il ne peut donc, à partir de là, ériger quelque nouveau système philosophique, mais seulement expliciter la philosophie que ce savoir véhicule, et le considérer moins pour en tirer une leçon philosophique que pour en saisir l'effectivité, fût-ce au point limite qu'est l'état momentané d'un savoir.

LE LANGAGE ENSEMBLISTE

Nous étudierons le langage ensembliste en nous appuyant le plus possible sur des notions intuitives. Cette façon de procéder permet d'acquérir très rapidement ce langage, indispensable pour progresser en mathématiques.

Par contre, elle présente des dangers : les notions intuitives, courantes, « naïves », sont trop imprécises. Ce sera l'objet du chapitre suivant traitant de la *théorie des ensembles* que de redéfinir plus rigoureusement certaines notions, de montrer sur quels choix elles reposent.

De toute façon, en mathématiques, il ne s'agit pas tant d'approfondir des notions « en soi » que d'étudier les rapports entre les divers êtres mathématiques. L'étude des relations existant entre plusieurs notions éclaire davantage chacune d'elles que leur étude isolée ou leur approfondissement à perte de vue.

Ensembles et éléments

Le mot **ensemble** est le mot le plus « neutre » que l'on puisse utiliser pour désigner un groupement imaginaire d'objets; mais il ne suffit pas d'imaginer un ensemble pour qu'il existe. Pourtant, dans un premier temps, nous nous en contenterons en nous limitant à des ensembles définis de façon indiscutable.

Ainsi : $\{0, 2, 4, 6, 8\}$ est un ensemble (l'ensemble des 5 premiers nombres pairs). De même : $\{x, y, z\}$ est l'ensemble de 3 objets x, y, z. Tout objet qui « fait partie » de l'ensemble considéré est appelé **« élément de l'ensemble »;** « x est élément de l'ensemble E » pourra s'écrire également : « $x \in E$ », qui se lit « x appartient à E » (E désigne à la fois l'ensemble et le nom de cet ensemble). Ainsi, si $E = \{0, 2, 4, 6, 8\}$, on peut écrire : $4 \in E$ et $0 \in E$. Par contre, 7 n'appartient pas à l'ensemble E : on notera $7 \notin E$, qui se lit « 7 n'appartient pas à E ». La notion d'appartenance n'est pas une notion de propriété : un même élément peut appartenir à plusieurs ensembles. Il n'y a pas de propriété privée en mathématiques !

On remarquera que, très souvent (mais pas toujours), les ensembles sont désignés par des lettres majuscules (E, A, B, ...) alors que les éléments sont désignés par des lettres minuscules (a, b, x, y, ...). Ceci se complique



Giraudon

▶ **La Mathématique** *d'Agostino di Duccio, sculpture ornant le bas-relief d'une chapelle en Italie (Rimini).*

du fait qu'un ensemble est un objet mathématique comme les autres; il peut donc être considéré comme élément d'un autre ensemble. Par exemple, une droite D est un ensemble de points; mais cette même droite est un élément de l'ensemble des droites d'un plan contenant D.

Les accolades sont réservées aux ensembles; elles entourent les éléments d'un ensemble. On distinguera par conséquent l'élément *a* (ou l'objet *a*) de *l'ensemble à un seul élément* $\{a\}$, appelé singleton; un ensemble à 2 éléments $\{a, b\}$ est une paire, ou doubleton.

Définition (construction) d'un ensemble

Un ensemble peut être défini de deux façons :

— par une description superficielle, mais complète, exhaustive, en citant tous ses éléments; c'est la définition en *extension.*

Exemple : $E = \{0, 2, 4, 6, 8\}$ est défini en extension, puisque l'on a cité tous ses éléments : 0, 2, 4, 6, 8.

Un objet qui n'est pas sur cette « liste » n'appartient pas à l'ensemble : $15 \notin E$, par exemple. Chaque élément de l'ensemble ne doit être écrit qu'une fois; si un même élément se retrouve écrit deux fois, il faut en supprimer une : $\{3, 7, 4, 7\} = \{3, 7, 4\}$; par contre, l'ordre n'a pas d'importance : $\{3, 7, 4\} = \{3, 4, 7\} = \{7, 4, 3\} = \ldots$

— par une description qui permette de comprendre quels sont les éléments qui appartiennent ou qui n'appartiennent pas à l'ensemble, sans les citer tous. C'est la définition en *compréhension,* par une propriété caractéristique des éléments de l'ensemble que l'on veut définir, c'est-à-dire par une propriété :

- que tous les éléments de l'ensemble ont;
- qu'ils sont les seuls à avoir.

Il est cependant nécessaire de « puiser » les éléments du nouvel ensemble dans un ensemble plus vaste (le référentiel), afin d'être sûr de définir un nouvel ensemble.

Exemple : si $\mathbb{N}$ désigne l'ensemble des entiers naturels ($\mathbb{N} = \{0, 1, 2, 3, \ldots\}$), nous pouvons définir l'ensemble E des nombres entiers naturels qui sont pairs et strictement inférieurs à 10 en écrivant :

$$E = \{x \in \mathbb{N} \mid x \text{ est pair et } x < 10\}.$$

Les accolades indiquent qu'il s'agit d'un ensemble; « $x \in \mathbb{N}$ » indique que les éléments de E sont des éléments de $\mathbb{N}$, c'est-à-dire des entiers naturels; la barre se lit « tels que » et « x est pair et $x < 10$ » est la propriété caractéristique des éléments de E. On lira donc : « E est égal à l'ensemble des x appartenant à $\mathbb{N}$ tels que x est pair et x est strictement inférieur à 10 ». Les entiers qui vérifient ces conditions sont : 0, 2, 4, 6, 8,

$$\text{donc } E = \{0, 2, 4, 6, 8\}.$$

Autre exemple : avec $\mathbb{R}$ = ensemble des nombres réels.

$$E = \{x \in \mathbb{R} \mid \sin x = 1\} = \left\{\frac{\pi}{2}, \left(\frac{\pi}{2} + 2\pi\right), \left(\frac{\pi}{2} + 4\pi\right), \ldots, \left(\frac{\pi}{2} + 2k\pi\right) \ldots\right\}$$

Nous insisterons enfin sur le fait qu'un ensemble n'est bien défini que lorsque l'on peut répondre par oui ou par non, de façon indiscutable, et pour tout objet, à la question : « Cet objet est-il élément de E ? »

Ainsi, dans l'exemple précédent : $15 \notin E$, $\pi \notin E$, $\frac{5\pi}{2} \in E$, la tour Eiffel $\notin$ E.

Par contre, l' « ensemble des grands champions automobiles », ou l' « ensemble des personnes à cheveux bruns » ne sont pas définis; ce ne sont pas des ensembles : les propriétés sont trop vagues, et se prêtent à des interprétations différentes pour un même individu.

Ensembles finis ou infinis

Un ensemble peut être fini ou infini :

$$\{0, 2, 4, 6, 8\} \text{ est fini.}$$

L'ensemble des nombres entiers pairs :

$$P = \{0, 2, 4, 6, 8, 10, 12, \ldots\} \text{ est infini.}$$

Ensemble vide

Un ensemble qui n'a pas d'éléments est un ensemble vide; on le note $\varnothing$ ou $\{\ \}$. Par exemple :

$$F = \{x \in \mathbb{N} \mid x > 7 \text{ et } x < 5\} = \varnothing$$

car aucun entier naturel n'est à la fois strictement supérieur à 7 et strictement inférieur à 5.

N = ensemble des nombres entiers naturels
= { 0, 1, 2, 3, ... }

Z = ensemble des nombres entiers relatifs
= { ..., − 3, − 2, − 1, 0, 1, 2, 3, ... }

D = ensemble des nombres décimaux, c'est-à-dire des nombres dont le développement décimal illimité ne contient que des zéros à partir d'un certain rang.
exemple : 7,25 = 7,250 00 ...
12 = 12,000 0 ...

Ils peuvent toujours se mettre sous la forme $\frac{x}{10^n}$, x étant un entier relatif.

Q = ensemble des nombres rationnels, c'est-à-dire des nombres x tels qu'il existe 2 entiers relatifs p et q vérifiant $x = \frac{p}{q}$.

Q est aussi l'ensemble des nombres dont le développement décimal illimité est *périodique* c'est-à-dire que le même «groupe de chiffres» se répète indéfiniment :

$\frac{11}{7} = 1{,}571\,428\,571\,428\,5 \ldots$ $\frac{13}{3} = 4{,}333\,3 \ldots$

La périodicité n'apparaît quelquefois qu'à la 30e décimale ou plus tard encore.

R = ensemble des nombres réels
= ensemble des nombres rationnels et irrationnels
= ensemble des nombres dont le développement décimal illimité est périodique ou apériodique (sans période)

nombres irrationnels : nombres dont le développement décimal illimité est *apériodique.*
$\sqrt{2}$ = 1,414 213 ...
π = 3,141 59 ...
sont des nombres irrationnels.

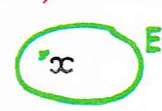
C = ensemble des nombres complexes

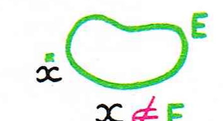
nombres complexes : nombres de la forme $a + ib$, où a et b sont deux réels et i un nombre «imaginaire pur» tel que $i^2 = -1$

on a : $\mathbb{N} \subset \mathbb{Z} \subset \mathbb{D} \subset \mathbb{Q} \subset \mathbb{R} \subset \mathbb{C}$

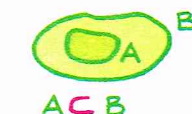
Richard Colin

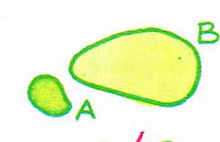
$\in$: relation d'appartenance entre un élément x et un ensemble E
$x \in E$ se lit «x appartient à E»
$x \notin E$ se lit «x n'appartient pas à E»





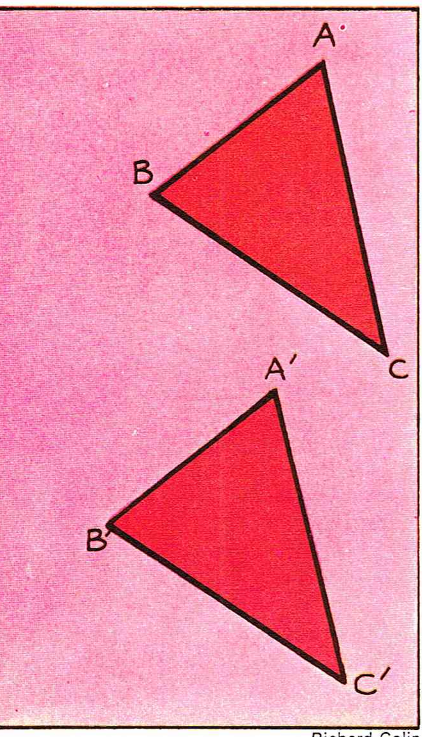
$\subset$: relation d'inclusion entre deux ensembles ou entre deux parties
$A \subset B$ se lit «A est inclus dans B»
$A \not\subset B$ se lit «A n'est pas inclus dans B»
$B \supset A$ se lit «B contient A»





$=$: relation d'égalité entre deux objets de même nature
$A = B \Rightarrow$ A et B ne sont qu'un seul et même objet

Richard Colin



Richard Colin

▲ *Les deux triangles* ABC *et* A'B'C' *se « ressemblent beaucoup »; ils sont même « superposables ». Mais ils ne sont pas égaux pour autant, car ils sont* distincts.

▶ *Ci-contre, de haut en bas, représentations schématiques, à l'aide de diagrammes de Venn, du* complémentaire d'un sous-ensemble; *de l'*intersection de deux ensembles; *de* deux ensembles disjoints.

Rien ne distingue un ensemble vide d'un autre ensemble vide; il n'y a donc qu'un ensemble vide (en fait, on pose l'existence d'un tel ensemble); et quel que soit l'objet x, $x \notin \varnothing$.

Égalité

La notion « moderne » d'égalité exclut la possibilité que deux objets distincts soient égaux. C'est en fait l'identité. Deux objets ne sont égaux que s'il s'agit d'un seul et même objet; $(3 + 2)$ et $(4 + 1)$ désignent en fait le même objet, le nombre 5; la relation d'égalité : $3 + 2 = 4 + 1$ est donc vérifiée. Par contre, deux triangles distincts ABC et A'B'C' (avec A différent de A' par exemple) ne peuvent être égaux, puisque distincts. C'est ce qui explique la disparition des défunts « cas d'égalité des triangles » des manuels scolaires récents; de même, on parlera d'aires égales et non de surfaces égales pour deux figures distinctes (l'aire étant la mesure de la surface), etc.

Ensembles égaux

Deux ensembles ne seront égaux que s'ils ont exactement les mêmes éléments, autrement dit que s'il ne s'agit que d'un seul et même ensemble.

Exemple :

$K = \{11, 13, 15, 17, 19\}$
$L = \{x \in \mathbb{N} \mid x \text{ est impair}, \ x > 10 \text{ et } x < 20\}$

L a pour éléments : 11, 13, 15, 17, 19
donc $K = L$.

Inclusion

Si l'on considère seulement certains éléments d'un ensemble, on considère une *partie* de cet ensemble; on dit encore un sous-ensemble de cet ensemble.

Exemple : $A = \{0, 4, 8\}$ est un sous-ensemble ou une partie de $E = \{0, 2, 4, 6, 8\}$ puisque A est constitué uniquement d'éléments de E. On note : $A \subset E$: « A est INCLUS dans E ». L'inclusion est une relation *entre deux ensembles* ou *sous-ensembles*.

Par définition :

$$A \subset E \Leftrightarrow [x \in A \Rightarrow x \in E]$$

Le signe $\Leftrightarrow$ se lit et signifie « équivalent à ».

Le signe $\Rightarrow$ se lit et signifie « implique » ou « entraîne ».

Si A est inclus dans E, tout élément de A est élément de E, et réciproquement, si tout élément de A est élément de E, alors A est inclus dans E. Il revient au même d'écrire $A \subset B$ (A est inclus dans B) ou $B \supset A$ (B contient A).

★ Tout ensemble est inclus dans lui-même. Pour tout ensemble E, $E \subset E$; E est la *partie pleine* de E.

★ L'ensemble vide $\varnothing$ est inclus dans tout ensemble : pour tout ensemble E, $\varnothing \subset E$; $\varnothing$ est la partie vide de E. (On peut en effet toujours affirmer que les éléments de $\varnothing$ sont aussi éléments de E; personne ne pourra montrer un élément de $\varnothing$ qui n'appartient pas à E, puisque $\varnothing$ n'a pas d'éléments !)

★ Toute partie de E qui n'est ni vide ni pleine est une *partie propre* de E.

★ Il existe un ensemble formé de toutes les parties de E, noté $\mathcal{P}$ (E).

Exemple : $A = \{x, y, z\}$.

$$\mathcal{P}(A) = \{\varnothing, \{x\}, \{y\}, \{z\}, \{x, y\}, \{y, z\}, \{z, x\}, \{x, y, z\}\}$$

Il y a donc 6 parties propres de A et $2^3 = 8$ parties de A au total.

Complémentaire d'un sous-ensemble

Soit E un ensemble, et A une partie de E. Les éléments de E qui ne sont pas éléments de A forment un ensemble, le complémentaire de A par rapport à E, noté $\complement_E A$ ou $(E - A)$ ou encore $\bar{A}$.

$$\complement_E A = \{x \in E \mid x \notin A\}$$ et :

$$x \in \complement_E A \Leftrightarrow x \in E \quad \text{et} \quad x \notin A$$

Exemple :

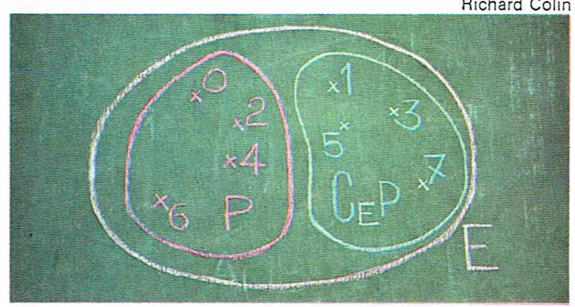
$E = \{0, 1, 2, 3, 4, 5, 6, 7\}$
$P = \{0, 2, 4, 6\}$

$\complement_E P = \{1, 3, 5, 7\}$

en particulier :

$$\complement_E \varnothing = E \qquad \complement_E E = \varnothing \qquad \complement_E(\complement_E A) = A$$

Richard Colin



« Opérations » sur les ensembles

Nous grouperons ici toutes les « opérations » qui associent à deux ensembles un troisième ensemble.

Intersection de deux ensembles

Soit A et B deux sous-ensembles d'un ensemble E; l'ensemble des éléments communs à A et à B, c'est-à-dire l'ensemble des éléments qui appartiennent à la fois à A et à B, est l'intersection de A et de B, notée « $A \cap B$ » et lue « A inter B » :

$$A \cap B = \{x \in E \mid x \in A \quad \text{et} \quad x \in B\}$$

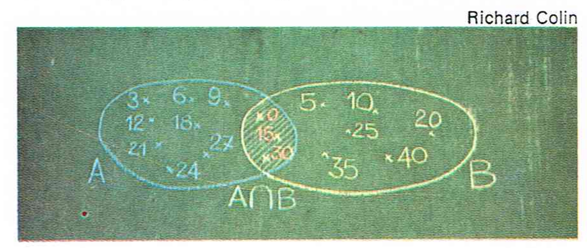
ou encore :

$$x \in A \cap B \Leftrightarrow x \in A \quad \text{et} \quad x \in B$$

Exemple :

$A = \{0, 3, 6, 9, 12, 15, 18, 21, 24, 27, 30\}$
$B = \{0, 5, 10, 15, 20, 25, 30, 35, 40\}$
$A \cap B = \{0, 15, 30\}$

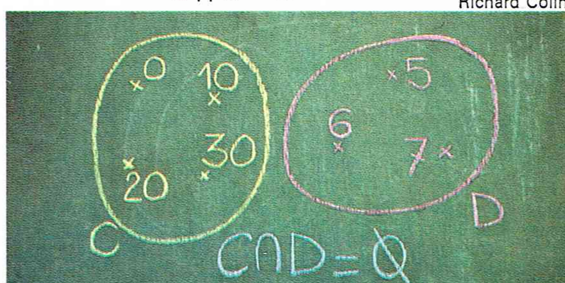
Richard Colin



L'intersection de deux ensembles peut être vide, mais elle existe toujours :

Exemple : $C = \{0, 10, 20, 30\}$
$D = \{5, 6, 7\}$
$C \cap D = \varnothing$

Richard Colin



$\cap$: intersection (de deux ensembles ou parties)
$A \cap B$ se lit : «A inter B»

$\cup$: union ou réunion (de deux ensembles ou parties)
$A \cup B$ se lit «A union B»

$\bigcap_{i \in I} A_i$: intersection de toutes les parties A_i dont l'indice i appartient à I (ensemble d'indices)
Si $I = \{1, 2, 3, ..., n\}$



$$\bigcap_{i \in I} A_i = A_1 \cap A_2 \cap A_3 \ldots \cap A_n$$

$\bigcup_{i \in I} A_i$: union de toutes les parties A_i telles que $i \in I$

A∩B

A∪B

Richard Colin

Les ensembles sont alors DISJOINTS.

Remarquons que, si $A \subset B$, alors $A \cap B = A$.

Exemple : $I = \{0, 1, 2, 3, 4, 5\}$
$J = \{0, 1, 2, 3, 4, 5, 6, 7, 8\}$
$I \subset J \Rightarrow I \cap J = I = \{0, 1, 2, 3, 4, 5\}$

et réciproquement : $I \cap J = I \Rightarrow I \subset J$.

Enfin, $\varnothing \cap A = \varnothing$ pour tout ensemble A.

Réunion de deux ensembles

Soit deux sous-ensembles A et B de E ; l'ensemble de tous les éléments de A et de B, c'est-à-dire l'ensemble des éléments qui appartiennent soit à A, soit à B, soit aux deux à la fois, s'appelle la réunion des ensembles A et B :

$$A \cup B = \{x \in E \mid x \in A \quad \text{ou} \quad x \in B\}$$

ou :

$$x \in A \cup B \Leftrightarrow x \in A \quad \text{ou} \quad x \in B$$

On remarquera que le « ou » utilisé ici est un « ou » inclusif, c'est-à-dire qu'il inclut la possibilité d'avoir les deux assertions à la fois ; par conséquent, les éléments de l'intersection sont aussi dans la réunion :

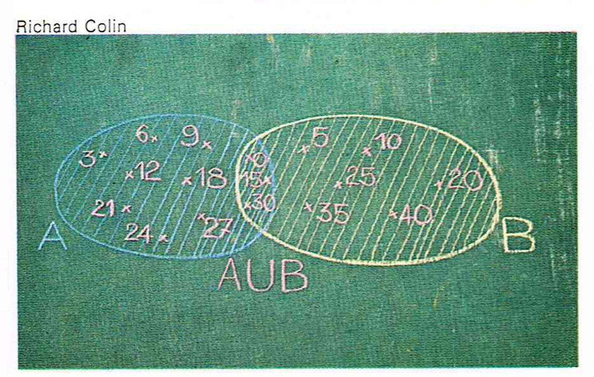
$$A \cap B \subset A \subset A \cup B$$
$$A \cap B \subset B \subset A \cup B$$

Remarquons que $A \cup A = A$ et $A \cup \varnothing = A$ pour tout ensemble A ; enfin, que, si $A \subset B$, alors $A \cup B = B$ et réciproquement : $A \cup B = B \Rightarrow A \subset B$.

Exemple : avec l'ensemble précédent (pris pour l'intersection) :

$A \cup B = \{0, 3, 6, 9\ 12, 15, 18, 21,$
$24, 27, 30, 5, 10, 20, 25, 35, 40\}$

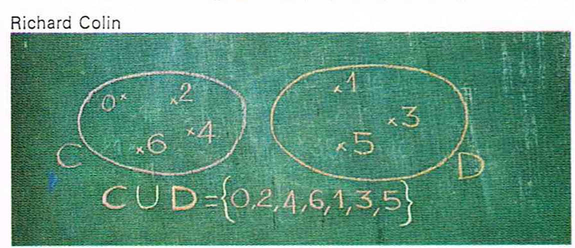
Richard Colin



◄ ***Représentation schématique, à l'aide d'un diagramme de Venn, de la* réunion *de deux ensembles.***

Autre exemple : $C = \{0, 2, 4, 6\}$
$D = \{1, 3, 5\}$
$C \cup D = \{0, 1, 2, 3, 4, 5, 6\}$

Richard Colin



◄ ***Un autre exemple de la* réunion *de deux ensembles.***

Différence symétrique de deux ensembles

Soit A et B deux sous-ensembles de E. La différence symétrique de A et de B est l'ensemble des éléments qui appartiennent soit à A seulement, soit à B seulement : on la note « $A \Delta B$ », qui se lit : « A delta B »

$$A \Delta B = \{x \in E \mid (x \in A \text{ et } x \notin B) \quad \text{ou} \quad (x \notin A \text{ et } x \in B)\}$$
$$= \{x \in E \mid x \in A \cup B \quad \text{et} \quad x \notin A \cap B\}$$

Plus simplement : $A \Delta B = \complement_{A \cup B} (A \cap B)$
$= (A \cup B) - (A \cap B)$.

Exemple : avec le même exemple que pour l'intersection et la réunion :

$A \Delta B = \{3, 6, 9, 12, 18, 21, 24, 27,$
$5, 10, 20, 25, 35, 40\}$

(on retire de $A \cup B$ tous les éléments de $A \cap B$).

Les propriétés de la réunion et de l'intersection seront étudiées dans le chapitre suivant.

Richard Colin

Δ	: différence symétrique de 2 ensembles $A \Delta B$ se lit : «A delta B» $A \Delta B = A \cup B - A \cap B$	A, B, $A \Delta B$
$\complement_E A$	: complémentaire de A dans E	A, E
$A \times B$	: produit cartésien de l'ensemble A par l'ensemble B. Se lit : «A croix B»	$A = \{a, b, c\}$ $B = \{x, y\}$ $A \times B = \{(a,x)(b,x)(c,x)(a,y)(b,y)(c,y)\}$
card A	: se lit : «Cardinal de A»	$A = \{a, b, c\}$ card A = 3

Relations binaires

Couples

Considérons une paire $\{x, y\}$; nous savons que $\{x, y\} = \{y, x\}$; autrement dit, l' « ordre » dans lequel sont rangés x et y n'a pas d'importance. Un couple sera, au contraire, un ensemble de deux éléments sur lequel a été défini un « ordre ».

Dans le couple (x, y), noté avec des parenthèses, « x est avant y », ou « x est la première projection (ou composante), y est la seconde projection (ou composante) ». Par conséquent, les couples (x, y) et (y, x) ne sont pas égaux.

Deux couples (x, y) et (x', y') sont égaux si, et seulement si, leurs premières projections sont égales entre elles et leurs secondes projections sont égales :

$$(x, y) = (x', y') \Leftrightarrow x = x' \text{ et } y = y'$$

Exemples :

$$(2 + 4, 10) = (6, 2 \times 5)$$
$$(15, 150) = (5 \times 3, 5 \times 30)$$

En fait, n'ayant pas vraiment défini ce qu'était un « ordre », la définition d'un couple repose sur une distinction formelle entre le premier et le second élément du couple, ainsi que nous le verrons au chapitre suivant.

Produit cartésien de deux ensembles

Soit E et F deux ensembles, et soit $E \times F$ (qui se lit « E croix F ») l'ensemble de *tous* les couples (x, y) tels que x est élément de E et y est élément de F :

$$E \times F = \{(x, y) \mid x \in E \quad \text{et} \quad y \in F\}$$

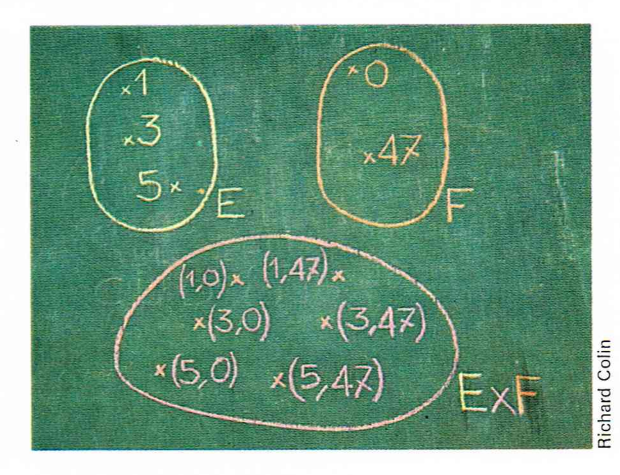
$E \times F$ est le produit cartésien de E par F.

Comme les couples (x, y) et (y, x) sont différents, les produits cartésiens $E \times F$ et $F \times E$ ne sont pas égaux. Il suffit d'inverser les couples de $E \times F$ pour obtenir $F \times E$.

On remarquera au passage que l'on n'a pas précisé d'ensemble de référence pour les couples (x, y) dans la définition de $E \times F$; c'est que l'on définit en fait le produit cartésien à partir de la définition rigoureuse d'un couple (*cf.* chapitre suivant).

Exemples :

— Soit $E = \{1, 3, 5\}$
$F = \{0, 47\}$
$E \times F = \{(1, 0), (1, 47), (3, 0), (3, 47), (5, 0), (5, 47)\}$
$F \times E = \{(0, 1), (0, 3), (0, 5), (47, 1), (47, 3), (47, 5)\}$



Richard Colin

◄ ***Diagramme de Venn du* produit cartésien *de deux ensembles.***

▼ **Plan muni du système de coordonnées** ***défini par R. Descartes (1596-1650), et à l'origine du terme*** **produit cartésien.**

— Soit un plan euclidien P muni d'un système d'axes de coordonnées cartésiennes. A chaque point M du plan sont alors associés deux nombres réels, son abscisse x et son ordonnée y, donc un couple (x, y), et réciproquement, à chaque couple (x, y) de deux nombres réels, est associé un point et un seul du plan.

On pourra identifier tout point M du plan P avec le couple (x, y) de ses coordonnées et raisonner dans $\mathbb{R} \times \mathbb{R}$, produit cartésien de $\mathbb{R}$ par $\mathbb{R}$ ($\mathbb{R}$ = ensemble des nombres réels), au lieu de raisonner dans le plan muni du système de coordonnées défini par René Descartes (1596-1650) : c'est là l'origine du terme produit cartésien.

Richard Colin

On note E^2 le produit cartésien de E par lui-même, et plus généralement, E^n le produit cartésien de n fois l'ensemble E.

Exemple : $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$; $\mathbb{R}^4 = \mathbb{R} \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}$

$$\mathbb{R}^n = \underbrace{\mathbb{R} \times \mathbb{R} \times \ldots \times \mathbb{R}}_{n \text{ fois}}$$

Un élément de $\mathbb{R}^4$ est un quadruplet : (x_1, x_2, x_3, x_4) avec $x_1 \in \mathbb{R}$, $x_2 \in \mathbb{R}$, $x_3 \in \mathbb{R}$ et $x_4 \in \mathbb{R}$.

Un élément de $\mathbb{R}^n$ est un *« n-uplet » :* $(x_1, x_2, \ldots, x_n)$ avec $x_1 \in \mathbb{R}, \ldots, x_n \in \mathbb{R}$.

Diagonale de E^2

Soit E un ensemble, et $E \times E = E^2$ le produit cartésien de E par lui-même. La diagonale de E^2 est l'ensemble de tous les couples (x, x) avec $x \in E$. On la note souvent Δ (delta) ou Δ_E :

$$\Delta_E = \{(x, y) \in E \times E \mid x = y\}$$

(Naturellement cette partie de $E \times E$ ne peut être définie dans un produit $E \times F$ avec $F \neq E$.)

Graphe d'une relation

Soit E et F deux ensembles, et $E \times F$ le produit cartésien de E par F. Une relation binaire de E vers F, c'est au fond l'ensemble des « liens » qui existent entre un élément de E et un élément de F (d'où le nom de binaire, puisqu'elle met en cause *deux* éléments à la fois) ; elle est donc définie par un ensemble de couples (x, y), tels que x est lié à y, avec $x \in E$ et $y \in F$, c'est-à-dire par une partie de $E \times F$. Cette partie du produit cartésien $E \times F$ est le graphe de la relation.

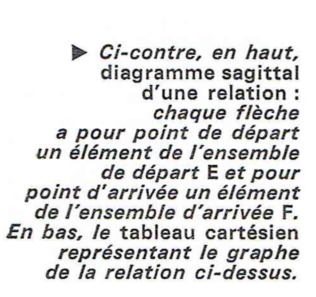
Réciproquement, toute partie de $E \times F$ définit une relation binaire de E vers F (puisqu'elle définit un ensemble de couples (x, y), donc des « liens »).

E est l'ensemble de départ et F est l'ensemble d'arrivée (puisque l'on va de E vers F).

Les relations binaires sont souvent désignées par des majuscules anglaises $\mathcal{R}$, $\mathcal{S}$, $\mathcal{T}$ ou par des symboles (qui désignent souvent une « loi de composition interne ») : $\top$, $\perp$, $\star$, Δ, etc. ($\top$ se lit souvent « truc » et $\perp$ « anti-truc »).

Si, par exemple, G est le graphe de la relation $\mathcal{R}$ de E vers F, pour exprimer que x (élément de E) est lié par $\mathcal{R}$ à y (élément de F), on peut écrire :

« $(x, y) \in G$ » ou « $x \mathcal{R} y$ » ou « $\mathcal{R}(x, y)$ »


▶ *Ci-contre, en haut,* **diagramme sagittal d'une relation :** *chaque flèche a pour point de départ un élément de l'ensemble de départ* **E** *et pour point d'arrivée un élément de l'ensemble d'arrivée* **F.** *En bas,* **le tableau cartésien** *représentant le graphe de la relation ci-dessus.*

Si x n'est pas lié par $\mathcal{R}$ à y, on écrira : $(x, y) \notin G$ ou « x non $\mathcal{R}$ y » ou « non $\mathcal{R}$ (x, y) ».

Une relation entre deux objets est une affirmation, une assertion, qui peut être ou ne pas être vérifiée. On admettra que, si une relation entre deux objets X et Y n'est pas vérifiée, son « contraire » est vérifié.

Par exemple :

si $a \in E$ n'est pas vrai, $a \notin E$ est vrai
si $A \subset B$ n'est pas vrai, $A \not\subset B$ est vrai
(A non inclus dans B)

Exemples :

— Soit $E = \{3, 4, 5\}$
$F = \{9, 12, 17\}$
$E \times F = \{(3, 9), (3, 12), (3, 17), (4, 9), (4, 12), (4, 17), (5, 9), (5, 12), (5, 17)\}$

Soit $G = \{(3, 9), (3, 12), (4, 12)\}$ le graphe de la relation. G définit en fait la relation $\mathcal{R}$: « ... divise ... » (au sens de la division euclidienne habituelle de reste nul) de E vers F.

En effet : 3 divise 9, 3 divise 12 et 4 divise 12, mais 3 ne divise pas 17, 4 ne divise pas 9, etc., donc

$$(3, 17) \notin G \,; \quad (4, 9) \notin G, \text{ etc.}$$

Le graphe G d'une relation binaire de E vers F est donc bien l'ensemble des couples (x, y) de $E \times F$ tels que x est lié à y par la relation :

$$G = \{(x, y) \in E \times F \mid x \mathcal{R} y\}$$

— Soit E un ensemble, $\mathcal{P}(E)$ l'ensemble de ses parties. L'inclusion entre deux parties de E est une relation de $\mathcal{P}(E)$ vers lui-même. Le graphe de cette relation est l'ensemble des couples (A, B) de parties de E tels que $A \subset B$.

Diagramme sagittal d'une relation

C'est le diagramme « avec les flèches », à partir des diagrammes de Venn des ensembles. Chaque flèche a pour point de départ un élément de l'ensemble de départ et pour point d'arrivée l'élément de l'ensemble d'arrivée lié à l'élément de départ considéré.

Pour l'exemple précédent :

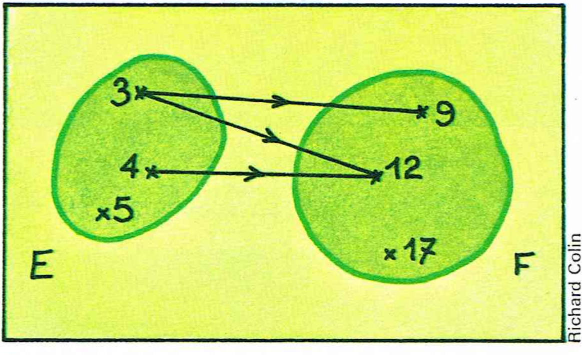
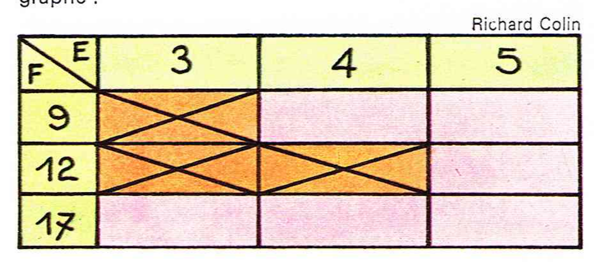


Tableau cartésien

Toujours avec l'exemple précédent, on met une croix dans chaque case correspondant à un élément du graphe :


Richard Colin

F \ E	3	4	5
9	X		
12	X	X	
17			

Ce tableau s'appelle « tableau cartésien » par analogie avec le plan muni d'un système de coordonnées cartésiennes. Remarquons que représenter le produit cartésien $E \times F$ dans ce tableau revient à mettre une croix dans chaque case.

Relation binaire dans un ensemble

Lorsque l'ensemble d'arrivée est le même que l'ensemble de départ (E = F), on peut définir une relation $\mathcal{R}$ de E vers lui-même ou une « relation $\mathcal{R}$ dans E ».

Exemple :

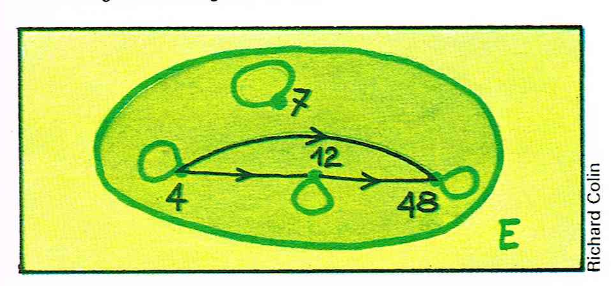
$$E = \{7, 4, 12, 48\}$$

$$E \times E = \{(7,7), (4,4), (12,12), (48,48), (7,4), (7,12), (7,48), (4,7), (4,12), (4,48), (12,7), (12,4), (12,48), (48,7), (48,4), (48,12)\}$$

$$G = \{(7,7), (4,4), (12,12), (48,48), (4,12), (12,48), (4,48)\}$$

G est le graphe de la relation « ... divise ... » définie dans E; mais l'on remarquera que les couples (7, 7) (4, 4), etc., ne peuvent faire partie du graphe que parce que E est à la fois ensemble de départ et d'arrivée.

Le diagramme sagittal est ici :



Richard Colin

Propriétés des relations binaires dans un ensemble

Soit E un ensemble et $\mathcal{R}$ une relation binaire définie dans E. La relation $\mathcal{R}$ définie dans E peut être :

★ ***Réflexive***

Pour tout $x \in E$, $x \mathcal{R} x$. Tout élément de E est lié à lui-même.

Exemple :

— l'égalité est réflexive : pour tout objet a, $a = a$;

— la relation de parallélisme dans l'ensemble des droites du plan est réflexive : toute droite est parallèle à elle-même (deux droites du plan sont parallèles si et seulement si elles sont d'intersection vide ou confondues).

★ ***Antiréflexive***

Pour tout $x \in E$, x non $\mathcal{R}$ x. Aucun élément de E n'est lié à lui-même.

Exemples :

— la relation < « strictement inférieur » définie dans $\mathbb{N}$ (ou dans $\mathbb{R}$) est antiréflexive : quel que soit le nombre x, x n'est pas strictement inférieur à x;

— la relation « D est perpendiculaire à D′ » dans l'ensemble des droites du plan est antiréflexive : aucune droite n'est perpendiculaire à elle-même.

★ ***Ni réflexive, ni antiréflexive***

Il existe $x \in E$, $y \in E$ tels que x non $\mathcal{R}$ x et $y \mathcal{R} y$. Certains éléments de E sont liés à eux-mêmes, d'autres non.

Exemple : la relation « ... divise ... » dans l'ensemble $H = \{0, 1, 2, 3\}$.

En effet, 1 divise 1, etc., mais 0 ne divise pas 0.

Dans les deux derniers cas, la relation est non réflexive.

★ ***Symétrique***

Pour tous $x \in E$, $y \in E$: $x \mathcal{R} y \Rightarrow y \mathcal{R} x$. Si x et y sont liés par $\mathcal{R}$, il en est de même pour y et x.

Exemple :

— l'égalité est symétrique pour tous objets x, y :

$$x = y \Rightarrow y = x$$

— la relation « D est parallèle à D′ » est symétrique :

$$D \mathbin{/\!/} D' \Rightarrow D' \mathbin{/\!/} D$$

— la relation « D est perpendiculaire à D′ » est symétrique : $D \perp D' \Rightarrow D' \perp D$.

★ ***Antisymétrique***

Pour tous $x \in E$, $y \in E$: $x \mathcal{R} y$ et $y \mathcal{R} x \Rightarrow x = y$.

Il n'existe pas deux éléments *distincts* appartenant à l'ensemble E qui soient liés « dans les deux sens » à la fois, c'est-à-dire que l'on ait à la fois $x \mathcal{R} y$ et $y \mathcal{R} x$.

Exemples :

— la relation $\leqslant$ dans $\mathbb{N}$

$$3 \leqslant x \text{ et } x \leqslant 3 \Rightarrow x = 3$$

— la relation d'inclusion dans $\mathcal{P}$ (E), ensemble des parties de E.

★ ***Ni symétrique ni antisymétrique***

Il existe x, y, z, t, éléments de E tels que :

$$x \mathcal{R} y \text{ et } y \mathcal{R} x$$
$$\text{et } z \mathcal{R} t \text{ et } t \text{ non } \mathcal{R} z$$

Il existe des couples liés symétriquement et d'autres qui ne le sont pas.

Dans les deux derniers cas, la relation est non symétrique.

★ ***Transitive***

Pour tous $x \in E$, $y \in E$, $z \in E$

[ou pour tout $(x, y, z) \in E \times E \times E$] :

$$x \mathcal{R} y \text{ et } y \mathcal{R} z \Rightarrow x \mathcal{R} z$$

Exemples :

— l'égalité est transitive pour tous les objets a, b, c : $a = b$ et $b = c \Rightarrow a = c$;

— la relation $\leqslant$ est transitive :

$$3 \leqslant 5 \text{ et } 5 \leqslant 7 \Rightarrow 3 \leqslant 7$$

★ ***Non transitive***

Il existe $x \in E$, $y \in E$, $z \in E$ tels que :

$$x \mathcal{R} y \text{ et } y \mathcal{R} z \text{ et } x \text{ non } \mathcal{R} y$$

Exemple : dans un ensemble $\{x, y, z, t\}$ « rangé » ou ordonné, la relation :

y est le successeur immédiat (ou le « voisin ») de x

z est le successeur immédiat (ou le « voisin ») de y

n'entraîne pas que z soit le successeur immédiat de x.

• • • •
x y 3 t

Relation d'équivalence

Une relation binaire définie dans un ensemble E qui est réflexive, symétrique et transitive est une relation d'équivalence dans E.

Exemples :

— l'égalité : $x = y \Leftrightarrow x$ et y sont un seul et même objet. L'égalité sur un ensemble E est une relation d'équivalence dont le graphe est l'ensemble des couples (x, x), donc la diagonale de $E \times E$;

— soit $\mathcal{R}$ la relation définie dans $\mathbb{N} \times \mathbb{N}$ par :

$$(x, y) \mathcal{R} (x', y') \Leftrightarrow [x - y = x' - y' \text{ ou } y - x = y' - x']$$

$\mathcal{R}$ est une relation d'équivalence (deux couples d'entiers sont équivalents s'ils ont la même différence entre leurs projections).

D'une façon générale, les relations, définies sur un ensemble, fondées sur « même », sont des relations d'équivalence.

Par exemple, dans $\mathcal{P}$ (E), ensemble des parties de E, la relation « A a le même cardinal que B », avec A et B éléments de $\mathcal{P}$ (E) (donc A et B parties de E), est une relation d'équivalence. En effet : A a le même cardinal que A; si A et B ont le même cardinal, B et A ont aussi le même cardinal; si A et B ont le même cardinal et si B et C ont le même cardinal, alors A et C ont le même cardinal; et ceci pour toutes parties A, B, C de E (le nombre d'éléments d'un ensemble est le cardinal de cet ensemble).

Éléments équivalents

Soit $\mathcal{R}$ une relation d'équivalence sur E. Deux éléments x et y de E liés par $\mathcal{R}$ sont dits « équivalents suivant la relation $\mathcal{R}$ » ou « équivalents modulo $\mathcal{R}$ ».

Dans certains cas, on utilise la relation d'équivalence notée : $x \equiv y$ (mod. a), qui se lit : « x congru à y modulo a ». C'est la congruence modulo a, qui exprime que deux éléments x et y d'un ensemble répondant par ailleurs à certaines conditions (c'est un groupe commutatif noté additivement) sont liés si $x - y$ est un multiple de a (il existe q entier, positif ou négatif, tel que $x - y = qa$).

Exemple : dans $\mathbb{Z}$ muni de l'addition : $85 \equiv 65$ (mod. 4), car $85 - 65 = 5 \times 4$, de même $16 \equiv 12$ (mod. 2).

($\mathbb{Z}$ est l'ensemble des entiers relatifs, c'est-à-dire des entiers positifs ou négatifs.)

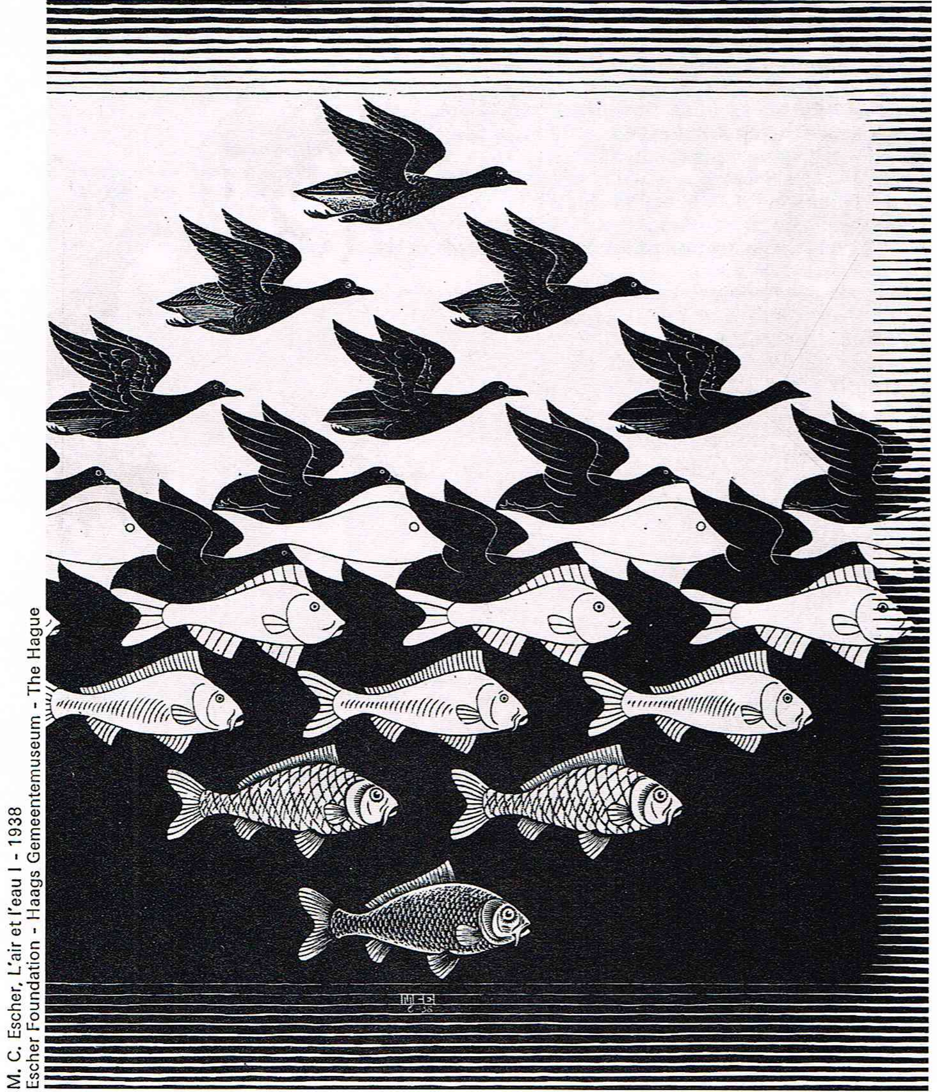
Classe d'équivalence modulo $\mathcal{R}$, ou suivant la relation $\mathcal{R}$

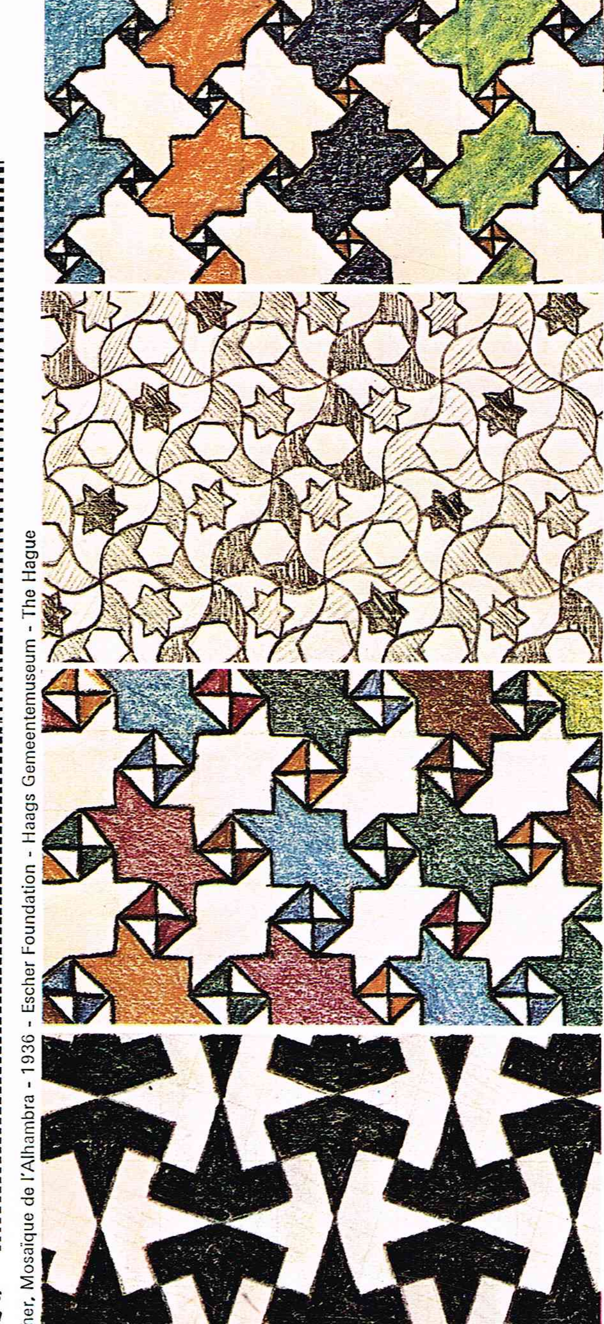
Soit x un élément de l'ensemble E. L'ensemble des éléments y de E liés à x, donc équivalents à x, est « la classe d'équivalence de x modulo $\mathcal{R}$ », ou « la classe d'équivalence de x suivant la relation $\mathcal{R}$ » :

$$C_x = \{y \in E \mid x \mathcal{R} y\}$$

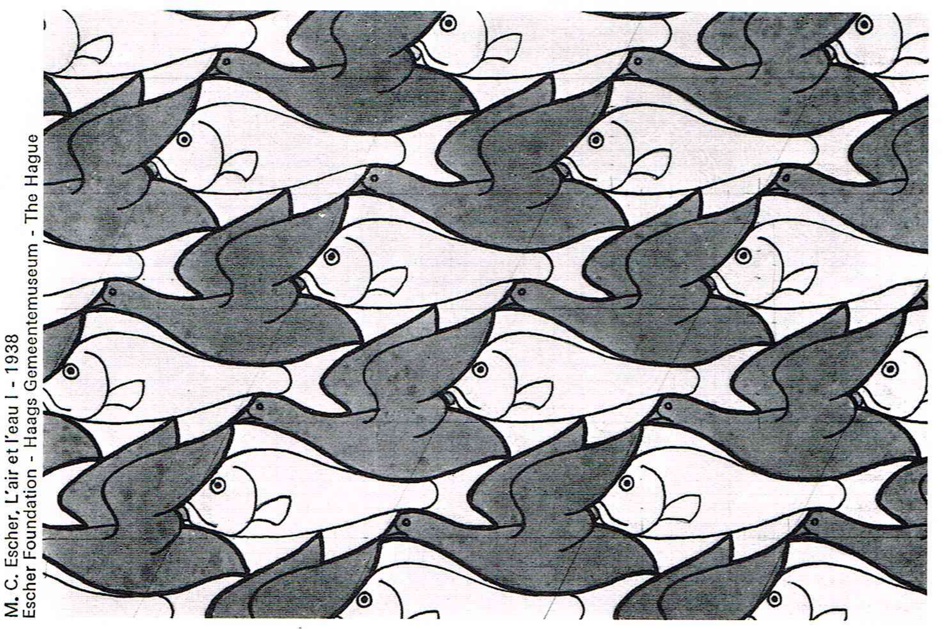
On note également $\dot{x}$ la classe d'équivalence de x.

Tout élément d'une classe d'équivalence est un représentant de cette classe.


M. C. Escher, L'air et l'eau I - 1938
Escher Foundation - Haags Gemeentemuseum - The Hague


M. C. Escher, Mosaique de l'Alhambra - 1936 - Escher Foundation - Haags Gemeentemuseum - The Hague

▲ ▼ *A partir d'un « pavé » type, ou motif, il s'agit de recouvrir* **tout** *le plan, sans trou, en disposant régulièrement le même pavé. Ce motif est dessiné de façon à ce que le* **complémentaire** *de l'ensemble des poissons soit un* **ensemble** *d'oiseaux (naturellement, l'ensemble de ces motifs forme une partition du plan ; et l'ensemble des motifs, complets ou partiels, forme sur une page — une partie du plan — une* **partition** *de cette page).*


M. C. Escher, L'air et l'eau I - 1938
Escher Foundation - Haags Gemeentemuseum - The Hague

Si deux éléments sont équivalents, leurs classes d'équivalence sont égales :

$$x \mathcal{R} y \Leftrightarrow C_x = C_y \Leftrightarrow y \in C_x \Leftrightarrow x \in C_y$$

Partition d'un ensemble

Soit E un ensemble. Un ensemble $\{A_1, A_2, ..., A_n\}$ de parties de E forme une partition de E si et seulement si :

— Aucune partie n'est vide :

$$\text{pour tout } i : A_i \neq \varnothing \quad 1 \leqslant i \leqslant n$$

— L'intersection de deux parties distinctes est vide :

$$\text{pour tout } i, \ \text{pour tout } j, \ i \neq j : A_i \cap A_j = \varnothing \quad 1 \leqslant i \leqslant n \quad 1 \leqslant j \leqslant n$$

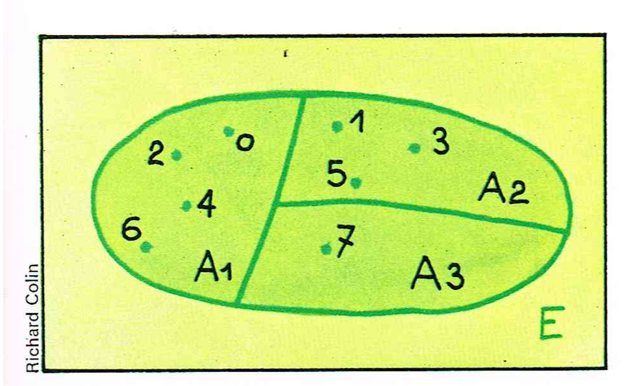
— La réunion de toutes les parties est égale à E.

$$\bigcup_{1 \leqslant i \leqslant n} A_i = E$$

Une assiette qui se casse en quelques morceaux fournit une bonne image d'une partition, souvent involontaire dans ce cas !

Exemple : $E = \{0, 1, 2, 3, 4, 5, 6, 7\}$
$A_1 = \{0, 2, 4, 6\}$
$A_2 = \{1, 3, 5\}$
$A_3 = \{7\}$

$\{A_1, A_2, A_3\}$ est une partition de E.



◀ *Représentation schématique de la* **partition d'un ensemble E.**

Partition de E en classes d'équivalence

L'ensemble des classes d'équivalence suivant une relation $\mathcal{R}$ définie sur E est une partition de E.

Exemple : soit $E = \{a, b, c\}$

$$\mathcal{P}(E) = \{ \varnothing, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\} \}$$

Soit $\mathcal{R}$ la relation d'équivalence définie sur E par : « ... a le même cardinal que ... »

Les classes d'équivalence suivant $\mathcal{R}$ sont :

$A_1 = \{\varnothing\}$: ensemble des parties à 0 élément ;

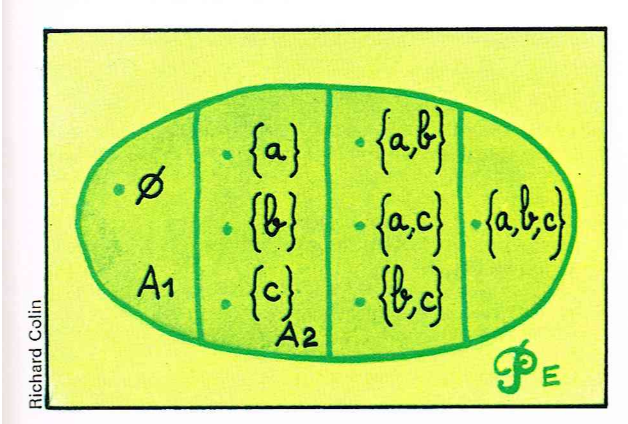
$A_2 = \{\{a\}, \{b\}, \{c\}\}$: ensemble des parties à 1 élément ;

$A_3 = \{\{a, b\}, \{a, c\}, \{b, c\}\}$: ensemble des parties à 2 éléments ;

$A_4 = \{\{a, b, c\}\}$: ensemble des parties à 3 éléments.

On remarquera que $A_1 = \{\varnothing\}$ n'est pas vide, mais est un singleton, ensemble à un élément (ici cet élément est « l'objet » $\varnothing$).

Ces classes A_1, A_2, A_3, A_4 forment bien une partition de $\mathcal{P}(E)$: $\{A_1, A_2, A_3, A_4\}$.



◀ *Représentation schématique d'une* **partition de** $\mathcal{P}$ *(E) en* **classes d'équivalence.**

Ensemble quotient

L'ensemble des classes d'équivalence modulo la relation d'équivalence $\mathcal{R}$ définie sur l'ensemble E est l'ensemble quotient de E par la relation $\mathcal{R}$, noté $E/\mathcal{R}$.

Relations d'ordre dans un ensemble

Une relation définie dans un ensemble E qui est : réflexive, antisymétrique et transitive, est une *relation d'ordre large.*

Exemples :

— la relation $\leqslant$ (inférieur ou égal) dans $\mathbb{R}$ est une relation d'ordre large ;

— l'inclusion définie dans $\mathcal{P}(E)$ est une relation d'ordre large.

Une relation définie dans un ensemble E qui est : antiréflexive, antisymétrique et transitive, est une *relation d'ordre strict.*

Exemple : la relation $<$ (strictement inférieur) dans $\mathbb{R}$.

Une relation d'ordre (large) définit un ordre sur E : E muni de cette relation d'ordre, est un *ensemble ordonné.*

On note souvent $\leqslant$ ou $\preccurlyeq$ une relation d'ordre large, même s'il ne s'agit pas de relations d'inégalité habituelles (sur $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{R}$).

Une « relation d'ordre » (sans précisions) est une relation d'ordre large. Remarquons qu'une « relation d'ordre strict » n'est donc pas une relation d'ordre.

Ordre total, ordre partiel

L'ordre défini sur un ensemble E peut être :

★ ***Total :*** si deux éléments quelconques de E peuvent être toujours classés ou comparés l'un par rapport à l'autre.

Pour tout $(x, y) \in E \times E$, $x \mathcal{R} y$ ou $y \mathcal{R} x$.

Exemple : la relation $\leqslant$ dans $\mathbb{R}$ définit un ordre total sur $\mathbb{R}$.

★ ***Partiel :*** il existe au moins deux éléments de E qui ne peuvent être classés ou comparés l'un par rapport à l'autre.

Il existe $(x, y) \in E \times E$: x non $\mathcal{R}\, y$ et y non $\mathcal{R}\, x$.

Exemples :

— L'inclusion dans $\mathcal{P}(E)$ définit un ordre partiel (par exemple, on ne peut classer par inclusion $\{a\}$ et $\{b, c\}$ ou $\{a, b\}$ et $\{a, c\}$) ;

— La relation « ... divise ... » définit dans $\mathbb{N}$ un ordre partiel (il existe par exemple [3, 20] tel que ni 3 ne divise 20, ni 20 ne divise 3).

Il y a donc deux possibilités pour un « ordre large » :
— ordre large et total
— ordre large et partiel

et deux possibilités pour un « ordre strict » :
— ordre strict et total
— ordre strict et partiel.

Fonction de E dans F

Soit $\mathcal{R}$ une relation de E vers F, de graphe G : $\mathcal{R}$ est une fonction si, et seulement si pour tout $(x, y) \in G$, et pour tout $(x', y') \in G$: $x = x' \Rightarrow y = y'$.

C'est-à-dire que tout élément de l'ensemble de départ est lié *au plus* à un élément de l'ensemble d'arrivée (donc soit à aucun élément, soit à un élément).

On note en général une fonction par une lettre minuscule : $f, g, h, \ldots$; une flèche ($\rightarrow$ ou $\mapsto$) relie les ensembles de départ et d'arrivée :

« $f : E \rightarrow F$ » ou « $f : E \mapsto F$ » ou « $E \xrightarrow{f} F$ »

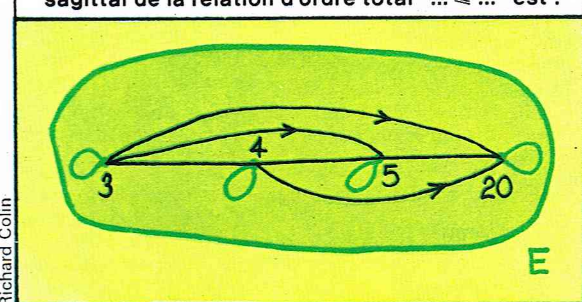
Exemples :

$$— f : \mathbb{R} \rightarrow \mathbb{R}, \quad x \rightarrow 3x + 2$$

$$— g : \mathbb{R} \rightarrow \mathbb{R}, \quad x \rightarrow \frac{3x}{x - 7}$$

$$— h : \mathbb{N} \rightarrow \mathbb{R}, \quad x \rightarrow \frac{3x}{x - 7}$$

Exemple : dans E = { 3, 4, 5, 20 }, le diagramme sagittal de la relation d'ordre total "... ⩽ ..." est :



◀ **Relation d'ordre total.**

Exemple : dans E = { 3, 4, 5, 20 }, le diagramme sagittal de la relation "... divise..." est :

3 ne divise ni 4, ni 5, ni 20 (le reste de la division de ces nombres par 3 n'est pas nul)

◀ **Relation d'ordre partiel.**

Domaine de définition d'une fonction

Soit f une fonction de E dans le graphe G. L'ensemble des éléments de E qui sont liés à un (et un seul) élément de F est le domaine de définition de la fonction f, noté D_f en général.

$$D_f = \{x \in E \mid \text{il existe } y : (x, y) \in G\}$$

Exemples : $f : \mathbb{R} \to \mathbb{R}$

$$x \to \frac{10x+2}{(5x+3)(x-5)}$$

On ne peut diviser un nombre par zéro; donc il faut que $(5x+3)$ et $(x-5)$ soient non nuls. Les valeurs qui annulent le dénominateur étant :

$-\frac{3}{5}$ et 5, le domaine de définition sera :

$$D_f = \mathbb{R} - \left\{-\frac{3}{5}, 5\right\} = \left]-\infty, -\frac{3}{5}\right[\cup \left]-\frac{3}{5}, 5\right[\cup \left]5, +\infty\right[$$

D_f est le complémentaire dans $\mathbb{R}$ de l'ensemble :

$$\left\{-\frac{3}{5}, 5\right\}.$$

Remarquons que si $g : \mathbb{N} \to \mathbb{R}$

$$x \to \frac{10x+2}{(5x+3)(x-5)}$$

$$D_g = \mathbb{N} - \{5\} = \complement_{\mathbb{N}}\{5\}$$

puisque 5 est le seul entier naturel qui annule le dénominateur, donc pour lequel l'expression $\frac{10x+2}{(5x+3)(x-5)}$ n'est pas définie.

Application

C'est une fonction dont le domaine de définition est égal à l'ensemble de départ tout entier. Tout élément x de l'ensemble de départ est donc lié à un et un seul élément y de l'ensemble d'arrivée; y est *l'image de x, notée f (x)* [qui se lit *f de x*] :

f application de E dans F, de graphe G (ou de graphe fonctionnel G) $\Leftrightarrow$ pour tout $x \in E$, il existe un et un seul élément $y \in F$ tel que $(x, y) \in G$.

Exemple : $f : \mathbb{R} \to \mathbb{R}$

$x \to ax + b$ avec $a, b \in \mathbb{R}$

L'image de x est : $f(x) = ax + b$

(si $f : x \to 3x + 2$, $f(5) = 17$)

En écrivant $y = f(x)$, c'est-à-dire $y = ax + b$, on retrouve l'équation d'une droite dans un système de coordonnées cartésiennes.

Image d'une partie A de E

Soit $f : E \to F$ une application et A une partie de E; l'ensemble des éléments de F qui sont l'image d'au moins un élément de A est l'image de A par f :

$$f(A) = \{y \in F \mid \text{il existe au moins un } x \in A \text{ tel que } y = f(x)\}$$

Exemple :

$f : \mathbb{R} \to \mathbb{R}$

$x \to 3x^2 + 1$

$f(5) = 3 \times 5^2 + 1 = 76$
$f(10) = 3 \times 10^2 + 1 = 301$
$f(1) = 3 \times 1^2 + 1 = 4$

Donc, si $A = \{1, 5, 10\}$ est une partie de $\mathbb{R}$, l'image de A par f est :

$$f(A) = \{f(1), f(5), f(10)\} = \{4, 76, 301\}$$

Image réciproque d'une partie de l'ensemble d'arrivée

Soit B une partie de F. L'ensemble de tous les éléments de E dont l'image est élément de B est l'image réciproque de B, notée $f^{-1}(B)$.

Exemple : avec $f : \mathbb{R} \to \mathbb{R}$

$x \to 3x^2 + 1$

Si $B = \{49, 76\}$

$f^{-1}(B) = \{-4, 4, -5, 5\}$

car $f(-4) = f(4) = 49$ et $f(-5) = f(5) = 76$.

Ensemble des applications de E dans F

L'ensemble de toutes les applications de E dans F est noté F^E, ou $\mathcal{F}(E, F)$. C'est une partie de l'ensemble $\mathfrak{P}(E \times F)$, puisque le graphe de toute application est une partie de $E \times F$.

Si $E = F$, on note cet ensemble E^E ou $\mathcal{F}(E)$.

Exemple : $\mathbb{R}^{\mathbb{R}} = \mathcal{F}(\mathbb{R})$ est l'ensemble de toutes les applications de $\mathbb{R}$ dans $\mathbb{R}$.

Application surjective

Si l'image de l'ensemble de départ, $f(E)$, est égale à l'ensemble d'arrivée F, l'application est surjective :

f application de E dans F :

f surjective $\Leftrightarrow f(E) = F$.

Exemple : $f : \mathbb{R} \to \mathbb{R}_+$ ($\mathbb{R}_+$ est l'ensemble des réels positifs).

$x \to x^2$

Application injective

Soit f une application de E dans F.

Si à deux éléments distincts x et x' correspondent deux images distinctes, l'application est injective :

f injective $\Leftrightarrow$ pour tout $(x, x') \in E \times E$:
$x \neq x' \Rightarrow f(x) \neq f'(x)$
$\Leftrightarrow$ pour tout $(x, x') \in E \times E$:
$f(x) = f(x') \Leftrightarrow x = x'$

(si deux éléments ont la même image, ils sont égaux).

Application bijective

C'est une application qui est à la fois injective et surjective.

Tout élément de E a alors une et une seule image, et tout élément de F est alors image d'un et d'un seul élément de E.

Composition de deux applications

Soit E, F, G, trois ensembles, et f, g deux applications :

$f : E \to F$ et $g : F \to G$

on peut également noter $E \xrightarrow{f} F$ et $F \xrightarrow{g} G$.

L'application composée « $g \circ f$ » qui se lit « g rond f » est une application de E dans G définie par :

$$\text{pour tout } x \in E : g \circ f(x) = g\{f(x)\}$$

Exemple :

$f : \mathbb{N} \to \mathbb{R}_+$
$x \to 3x + 2$
$g : \mathbb{R}_+ \to \mathbb{R}$
$x \to 5x^2 - 100$

$g \circ f : \mathbb{N} \to \mathbb{R}$
$x \to 5(3x + 2)^2 - 100.$

On ne peut ici définir $f \circ g$ car l'ensemble d'arrivée de g n'est pas l'ensemble de départ de f.

On peut écrire également : $g : \mathbb{R}_+ \to \mathbb{R}$
$y \to 5y^2 - 100$

et, en posant $y = f(x)$: $g \circ f(x) = g(f(x)) = g(y)$ dans l'exemple ci-dessus

$y = 3x + 2;$
$g \circ f(x) = g(y) = 5y^2 - 100 = 5(3x + 2)^2 - 100.$

Application identique dans un ensemble

C'est l'application qui, à tout élément de l'ensemble, associe ce même élément. On la note I_E, ou id_E

$I_E : E \to E$
$x \to x$

Son graphe est évidemment l'ensemble des couples (x, x) de $E \times E$, donc la diagonale de $E \times E$.

Application réciproque

Si $f : E \to F$ est une bijection, l'application réciproque $f^{-1} : F \to E$ est l'application telle que $g \circ f = f \circ g = I_E$.

Exemple : soit $f : \mathbb{R} \to \mathbb{R}$
$x \to 3x + 2.$

L'application réciproque est $f^{-1} : \mathbb{R} \to \mathbb{R}$

$$x \to \frac{1}{3}(x - 2)$$

Vérification :

$$f^{-1} \circ f(x) = \frac{1}{3}\left((3x + 2) - 2\right) = x = id_{\mathbb{R}}x$$

$$f \circ f^{-1}(x) = 3\left(\frac{1}{3}(x - 2)\right) + 2 = x = id_{\mathbb{R}}x$$



Richard Colin

▲ *Première notion* **d'ensemble : A,** *l'ensemble des formes jaunes ;* **B,** *l'ensemble des formes rouges.*

LES ENSEMBLES

La notion d'ensemble est une notion primitive des mathématiques ; nous ne donnerons donc pas de définition d'un ensemble, pas plus qu'en géométrie, par exemple, on ne définit ce qu'est un point. Cependant il ne faut pas en conclure que l'usage du mot *ensemble* est arbitraire, au contraire il est strictement régi par des axiomes, sans quoi l'on aboutirait rapidement à des contradictions.

En reprenant l'exposé des diverses notions introduites au chapitre précédent, nous allons dégager à chaque étape les *axiomes* nécessaires pour garantir l'existence cohérente d'objets mathématiques appelés ensembles. Une véritable présentation axiomatique de la théorie des ensembles reposant sur une bonne connaissance de la logique, le lecteur pourra se reporter à la bibliographie.

Dans une seconde partie, nous montrerons comment, à l'aide de méthodes reposant sur les propriétés des ensembles, on peut étendre les notions usuelles de *dénombrement* et de *numération*.

Les axiomes

Égalité - appartenance

Le principal concept (non défini) de la théorie des ensembles est celui d'**appartenance.** Si x appartient à A, on dira que x est un élément de A, et on écrira : $x \in A$; la négation étant notée $x \notin A$.

Une relation élémentaire entre ensembles est l'**égalité** $A = B$.

La notion d'appartenance est liée à celle d'égalité de la façon suivante : deux ensembles sont égaux si, et seulement si, ils ont les mêmes éléments **(axiome d'extensionalité).**

L'égalité entre ensembles est une relation :
- *réflexive :* $A = A$ pour tout ensemble A ;
- *symétrique :* si $A = B$, alors $B = A$;
- *transitive :* si $A = B$ et $B = C$, alors $A = C$.

Inclusion - sous-ensembles

Si A et B sont deux ensembles et si tout élément de A est élément de B, on dira que A **est inclus dans** B, ou que A est un **sous-ensemble** de B, ou que A est une **partie** de B ; on écrira indifféremment :

$$A \subset B \qquad B \supset A$$

L'inclusion entre deux ensembles est une relation :
- *réflexive :* $A \subset A$ pour tout ensemble A ;
- *transitive :* si $A \subset B$ et $B \subset C$, alors $A \subset C$;
- *antisymétrique :* si $A \subset B$ et $B \subset A$, alors $A = B$.

La plupart des démonstrations d'égalité entre deux ensembles se font à l'aide de cette dernière propriété.

L'axiome de sélection - paradoxe de Russell

A part l'axiome d'extensionalité, tous les principes fondamentaux de la théorie des ensembles ont pour objet la construction de nouveaux ensembles à partir d'anciens. On le fera à l'aide de deux types d'énoncés :

$$x \in A \qquad A = B$$

par application des opérateurs logiques usuels et des règles sur ces opérateurs (voir *Logique*), mais il faudra imposer une restriction supplémentaire, sous peine de se heurter à des paradoxes.

En effet supposons que S (x) soit un énoncé du type précédent. On peut définir un ensemble pour

$$X = \{x \mid S(x)\}$$

X étant l'ensemble de tous les x pour lesquels la propriété S (x) est vraie
et en prenant pour S (x) l'énoncé $x \notin x$

$$X = \{x \mid x \notin x\}$$

on constate alors que la proposition $X \in X$ est antinomique car : — si $X \in X$ alors X élément de X vérifie la propriété caractéristique $X \notin X$;
— si $X \notin X$ la propriété S (X) est vraie donc $X \in X$; dans les deux cas on aboutit à une contradiction.

Ce paradoxe mis en évidence par Russell peut être illustré de façon plus concrète par l'exemple suivant : « Si le barbier d'un village rase tous les habitants qui ne se rasent pas eux-mêmes, ce barbier se rase-t-il lui-même ? S'il se rase il doit appartenir à la catégorie de ceux qui ne se rasent pas eux-mêmes, mais s'il ne se rase pas, il appartient au groupe de ceux qu'il doit raser. » Dans les deux cas, qui sont les seuls possibles, il y a contradiction.

▶ Page ci-contre, en haut, à gauche, Bertrand Russell, *mathématicien, philosophe et sociologue, est d'abord un logicien, l'un des promoteurs de la logistique. On lui doit notamment* Principia mathematica *qu'il rédigea en collaboration avec N. Whitehead en 1903.*

On élimine ce paradoxe en posant le principe fondamental suivant : à tout ensemble A et toute condition S (x), il correspond un ensemble B dont les éléments sont exactement les éléments de A pour lesquels S (x) est vraie **(axiome de sélection).**

Ainsi, pour tout ensemble A, on peut construire l'ensemble B tel que :

$$B = \{x \in A \mid S(x)\}$$

Les éléments de B sont tous les éléments de A pour lesquels S (x) est vraie.

En particulier : $B = \{x \in A \mid x \notin x\}$ est un ensemble, appelé partie russellienne de A. Cet ensemble est tel que $B \notin A$.

En effet, si $B \in B$, alors B vérifie les conditions imposées aux éléments de B, à savoir $B \in A$ et $B \notin B$; si $B \notin B$, alors on a $B \notin A$ ou $B \in B$. Seule la deuxième hypothèse n'aboutit pas à une contradiction, mais à condition que $B \notin A$.

L'ensemble A est arbitraire, on voit ainsi qu'il existe toujours quelque chose (B) qui n'appartient pas à un ensemble quelconque A.

En d'autres termes, rien ne contient tout. De même, on démontre que l'ensemble de tous les ensembles n'existe pas.

Nous savons maintenant qu'il est possible de construire un nouvel ensemble en donnant une propriété caractéristique de certains éléments d'un ensemble déjà connu. C'est ce qu'on appelle la définition *en compréhension.*

Ensembles particuliers

Faisons l'hypothèse qu'il existe un ensemble A, ce qui sera justifié ultérieurement (on posera comme axiome l'existence d'ensembles infinis). On peut construire :

$$\{x \in A \mid x \neq x\}$$

cet ensemble n'a aucun élément, c'est l'**ensemble vide,** noté $\varnothing$. On démontre que l'ensemble vide est sous-ensemble de tout ensemble.

Si a et b sont deux ensembles, il existe un ensemble A tel que $a \in A$ et $b \in A$, c'est ce qu'affirme l'**axiome de la paire.**

L'ensemble $\{x \in A \mid x = a \quad \text{ou} \quad x = b\}$ qui ne contient que a et b est noté $\{a, b\}$.

Si $b = a$ alors $\{a, b\} = \{a, a\} = \{a\}$

a est le seul élément de l'ensemble $\{a\}$.

Ainsi, on peut construire l'ensemble $\{\varnothing\}$ qu'il ne faut pas confondre avec $\varnothing$, car ce dernier ensemble n'a pas d'élément, alors que $\{\varnothing\}$ en a un : $\varnothing$.

Unions et intersections

C'est un nouvel axiome qui va nous permettre de réunir les éléments d'une famille d'ensembles en un ensemble global :

pour toute famille d'ensembles $\mathcal{F}$, il existe un ensemble qui contient tous les éléments appartenant à au moins un ensemble de la famille **(axiome de la réunion).**

On note cet ensemble : $\bigcup \mathcal{F}$ ou $\bigcup_{X \in \mathcal{F}} X$.

Dans le cas particulier d'une paire $\{A, B\}$ on adopte la notation $A \cup B$. De la définition de la réunion il résulte que x appartient à $A \cup B$ si, et seulement si x appartient soit à A, soit à B, soit aux deux :

$$A \cup B = \{x \mid x \in A \quad \text{ou} \quad x \in B\}$$

Les axiomes d'extensionalité et de sélection suffisent à définir et à assurer l'unicité de l'**intersection** d'une famille non vide d'ensembles.

Comme $\mathcal{F} \neq \varnothing$, on peut considérer un élément quelconque A de $\mathcal{F}$, alors :

$$\{x \in A \mid x \in X \text{ pour chaque X de } \mathcal{F}\}$$

définit un ensemble qui, en fait, ne dépend pas du choix arbitraire de A. On note cet ensemble :

$$\bigcap \mathcal{F} \quad \text{ou} \quad \bigcap_{X \in \mathcal{F}} X$$

et dans le cas d'une paire

$$A \cap B = \{x \in A \mid x \in B\} = \{x \in B \mid x \in A\} = \{x \mid x \in A \text{ et } x \in B\}.$$

Les propriétés de la *réunion* et de l'*intersection* sont les suivantes :

$A \cup \varnothing = A$

$A \cap \varnothing = \varnothing$

$A \cap A = A$ }
$A \cup A = A$ } *idempotence*

$A \cup B = B \cup A$ }
$A \cap B = B \cap A$ } *commutativité*

$A \cup (B \cup C) = (A \cup B) \cup C$ }
$A \cap (B \cap C) = (A \cap B) \cap C$ } *associativité*

$A \cap (A \cup B) = A$ }
$A \cup (A \cap B) = A$ } *absorption*

$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ }
$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ } *distributivité*

$A \subset B$ si et seulement si $A \cup B = B$

$A \subset B$ si et seulement si $A \cap B = A$.

Complémentaires et parties

Si A et B sont deux ensembles, la différence entre A et B, ou **complémentaire de** B **dans** A, est l'ensemble défini par :

$$A - B = \{x \in A \mid x \notin B\} = \complement_A B$$

Il n'est pas nécessaire de supposer $B \subset A$.

Si A et B sont des sous-ensembles d'un même ensemble E, on posera $\complement_E A = \overline{A}$ et $\complement_E B = \overline{B}$.

Les propriétés fondamentales de la *complémentation* sont données par :

$A \cap \overline{A} = \varnothing$

$A \cup \overline{A} = E$

$\overline{A \cap B} = \overline{A} \cup \overline{B}$ }
$\overline{A \cup B} = \overline{A} \cap \overline{B}$ } *lois de Morgan*

$A \subset B$ si et seulement si $A \cap \overline{B} = \varnothing$

$A \subset B$ si et seulement si $\overline{A} \cup B = E$

$A \subset B$ si et seulement si $\overline{B} \subset \overline{A}$.

On appelle différence symétrique de A et de B l'ensemble $A \Delta B = (A - B) \cup (B - A)$.

Cette opération est *commutative, associative* et est telle que $A \Delta \varnothing = A$ et $A \Delta A = \varnothing$.

La collection de tous les sous-ensembles d'un ensemble donné E est encore un ensemble, c'est ce qu'assure l'**axiome des parties** d'un ensemble :

$\mathfrak{P}(E) = \{X \mid X \subset E\}$ ensemble des parties de l'ensemble E.

Ainsi :

$$\mathfrak{P}(\{a, b\}) = \{\varnothing, \{a\}, \{b\}, \{a, b\}\}$$

$$\mathfrak{P}(\varnothing) = \{\varnothing\}, \qquad \mathfrak{P}(\{\varnothing\}) = \{\varnothing, \{\varnothing\}\}, \text{ etc.}$$

Définition d'un ensemble

Définition en extension

$E = \{1, 2, 3, 4, 6, 12\}$

On dresse la liste des éléments de l'ensemble considéré.

Définition en compréhension.

$E = \{x \mid x \in \mathbb{N}$ et x divise $12\}$

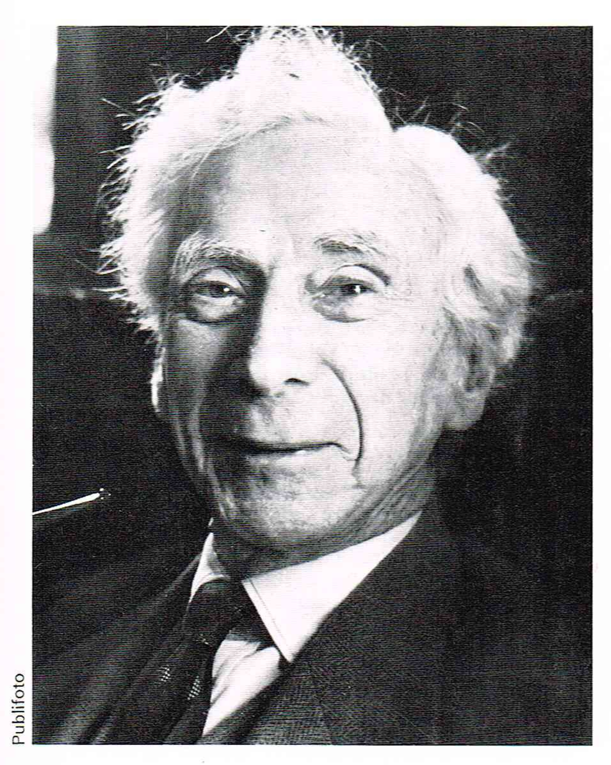
se lit « ensemble des x tel que x est un entier naturel et x divise 12 ».

Plus généralement :

$E = \{x \in A \mid S(x)\}$

« les éléments de E sont tous les éléments de A pour lesquels la propriété S (x) est vraie. »

Richard Colin


Publifoto

Couple - produit cartésien

La notion de paire $\{a, b\}$ va être complétée par celle de paire ordonnée ou couple.

Par définition, le **couple** (x, y) est l'ensemble $\{x, \{x, y\}\}$, définition due à Kuratowski.

x est le premier terme (première projection) du couple, y le second terme (seconde projection).

Alors que $\{a, b\} = \{b, a\}$ en général (si $a \neq b$) $(a, b) \neq (b, a)$.

On démontre que deux couples sont égaux si, et seulement si leurs projections sont égales.

Soit deux ensembles A et B, existe-t-il un ensemble qui contienne tous les couples (a, b) avec a dans A et b dans B ?

Comme $a \in A$, $\{a\} \subset A$; de même $\{b\} \subset B$, donc $\{a, b\} \subset A \cup B$. Par suite $\{a\} \in \mathcal{P}(A \cup B)$; $\{a, b\} \in \mathcal{P}(A \cup B)$ et la paire $\{\{a\}, \{a, b\}\}$ est un sous-ensemble de $\mathcal{P}(A \cup B)$, donc appartient à $\mathcal{P}(\mathcal{P}(A \cup B))$.

Tous les couples (a, b) appartiennent à l'ensemble $\mathcal{P}(\mathcal{P}(A \cup B))$ dès que $a \in A$ et $b \in B$. L'axiome d'extensionalité permet d'assurer l'existence d'un ensemble que l'on notera $A \times B$ tel que :

$$A \times B = \{x \in \mathcal{P}(\mathcal{P}(A \cup B)) \mid x = (a, b), a \in A, b \in B\}$$

l'axiome de sélection en garantit l'unicité.

$A \times B = \{(a, b) \mid a \in A, b \in B\}$ est le **produit cartésien** des ensembles A et B.

Relations binaires

Étant donné deux ensembles A et B, on appelle **graphe** ou **relation de A vers B** tout sous-ensemble $\mathcal{R}$ du produit cartésien $A \times B$.

A est l'ensemble de départ, B l'ensemble d'arrivée. Si $(x, y) \in \mathcal{R}$, on dit que x est un antécédent de y, y est une image de x par $\mathcal{R}$. Au lieu de noter $(x, x) \in \mathcal{R}$, on pose souvent $x \mathcal{R} x$.

Si $A = B$, alors tout sous-ensemble de $A \times A$ définit une relation binaire sur A.

Ainsi la relation d'inclusion entre les sous-ensembles d'un ensemble U est caractérisée par le sous-ensemble de $\mathcal{P}(U) \times \mathcal{P}(U)$ formé des couples (A, B) tels que $A \subset B$.

Préordre et ordre

Une relation binaire dans un ensemble E est une **relation de préordre** (ou un **préordre**) si elle est *réflexive* et *transitive*, c'est-à-dire vérifie :

— pour tout x, $x \mathcal{R} x$;
— si $x \mathcal{R} y$ et $y \mathcal{R} z$ alors $x \mathcal{R} z$.

Richard Colin

▲ **Intersection :** ***l'ensemble A des enfants qui portent au moins un vêtement rouge et l'ensemble B des enfants qui portent au moins un vêtement bleu se coupent ; l'enfant habillé de rouge et de bleu correspond à l'intersection des ensembles A et B.***



▲ **Inclusion :** ***l'ensemble des triangles jaunes est inclus dans l'ensemble des formes de couleur jaune.***

Richard Colin

Une relation binaire dans un ensemble E est une **relation d'ordre** (ou un **ordre**) si elle est *réflexive, antisymétrique* et *transitive,* soit :

— pour tout x, $x \mathcal{R} x$
— si $x \mathcal{R} y$ et $y \mathcal{R} x$, alors $x = y$
— si $x \mathcal{R} y$ et $y \mathcal{R} z$, alors $x \mathcal{R} z$.

La relation d'inclusion dans $\mathcal{P}$ (E) est une relation d'ordre. Divers symboles sont employés pour traduire une relation d'ordre, citons $\leqslant$, $<$, $\subset$, $\subseteq$, $\precsim$, ...

La relation d'ordre est **totale** si deux éléments quelconques de l'ensemble sont comparables. La relation d'inclusion dans $\mathcal{P}$ (E) n'est pas un ordre total, c'est un ordre **partiel.** Un ensemble sur lequel on a défini une relation d'ordre est dit **ordonné.**

Soit (E, $\leqslant$) un ensemble ordonné et A une partie de E. L'élément m de E est un **majorant** de A si m est supérieur à tous les éléments de A :

$$\text{pour tout } x \in A \text{ on a } x \leqslant m.$$

Si μ est un élément de A tel que pour tout x de A, $x \leqslant \mu$, μ est le **plus grand élément** de A. Enfin on appelle **borne supérieure** de A le plus petit des majorants de A. De façon analogue, on définit les notions duales des précédentes : minorant, plus petit élément, borne inférieure.

Si toute partie non vide d'un ensemble ordonné admet un plus petit élément, l'ordre défini sur cet ensemble est un **bon ordre.**

Équivalence

On appelle **relation d'équivalence** ou **équivalence** sur un ensemble E tout préordre $\mathcal{R}$ symétrique sur E, c'est-à-dire une relation binaire sur E qui est :

— *réflexive :* pour tout x, $x \mathcal{R} x$;
— *transitive :* si $x \mathcal{R} y$ et $y \mathcal{R} z$, alors $x \mathcal{R} z$;
— *symétrique :* si $x \mathcal{R} y$, alors $y \mathcal{R} x$.

Lorsqu'il existe une relation d'équivalence sur E, l'image $\mathcal{R}$ (x) de x par $\mathcal{R}$ est appelée **classe d'équivalence** de x modulo $\mathcal{R}$; on la note généralement $\overline{x}$. $\mathcal{R}$ étant réflexive, $\overline{x}$ contient toujours un élément (lui-même), donc $\overline{x} \neq \varnothing$; les propriétés de $\mathcal{R}$ entraînent que si $x \mathcal{R} y$ alors $\overline{x} = \overline{y}$ et réciproquement si $\overline{y} = \overline{x}$, alors $x \mathcal{R} y$.

Pour connaître une classe d'équivalence, il suffit d'en citer un élément, ce sera le **représentant** de la classe.

L'ensemble des classes d'équivalence forme une **partition** de E, c'est-à-dire un recouvrement de E en sous-ensembles disjoints. Cet ensemble est l'ensemble quotient de E par la relation d'équivalence $\mathcal{R}$; on le note E/$\mathcal{R}$. On a E/$\mathcal{R} \subset \mathcal{P}$ (E).

Fonction - Application

Une relation de E vers F est une **fonction** si, pour tout couple (x, y) et tout couple (x', y') appartenant à $\mathcal{R}$, $x = x'$ implique $y = y'$ (condition d'univocité).

L'ensemble $\{x \in E \mid (x, y) \in \mathcal{R}\}$ est le domaine, ou encore le domaine de définition, de la fonction.

Une **application** de E vers F est une fonction dont le domaine de définition est E tout entier. On emploie souvent la notation $f : E \to F$.

L'ensemble de toutes les applications de E dans F est un sous-ensemble de l'ensemble des parties $\mathcal{P}$ (E $\times$ F) ; il sera noté F^E ou $\mathcal{F}$ (E, F).

Si A est un sous-ensemble de E, l'image de A par l'application f est l'ensemble des éléments y de F pour lesquels il existe au moins un x dans A tel que $f(x) = y$. On appelle image de f et on note Im f l'ensemble f (E). On désigne par f^{-1} (B) l'ensemble des éléments x de A pour lesquels $f(x)$ appartient à B (B étant un sous-ensemble de F).

Par définition, une application $f : E \to F$ est *surjective* quand $f(E) = F$. Elle est *injective* si $f(x) = f(x')$ implique $x = x'$. Elle est *bijective* quand elle est à la fois injective et surjective.

Si f est bijective, il y a correspondance univoque entre E et F, à tout x appartenant à E correspond un, et un seul, $y = f(x)$, alors on peut associer à tout y l'unique x ainsi défini. L'application $F \to E$ correspondante est dite *application réciproque* de f, on la note : f^{-1}.

On démontre qu'il existe une bijection entre l'ensemble $\mathcal{R}$ (E) des équivalences sur E et l'ensemble $\mathcal{P}$ (E) des partitions de E; il revient donc au même de se donner explicitement une équivalence sur E ou de se donner la partition correspondante.

Dénombrement

Nous allons maintenant présenter les principales notions attachées aux problèmes de **dénombrement** par l'intermédiaire des **cardinaux** et des **ordinaux.**

Dans le langage courant, un nombre cardinal est celui qui mesure la quantité (n objets), un nombre ordinal est celui qui indique l'ordre (n-ième objet) ; nous allons montrer comment ces notions peuvent s'étendre aux ensembles infinis.

Nous supposerons établie l'existence de $\mathbb{N}$, ensemble des entiers naturels (voir *Nombres*) dont la construction repose sur un nouvel axiome, l'**axiome de l'infini.** Nous admettrons l'axiome du choix (le produit cartésien d'une famille d'ensembles non vides est non vide ou, ce qui est équivalent, pour tout ensemble E il existe une application f de $\mathfrak{P}$ (E) — $\varnothing$ dans E qui à tout X non vide inclus dans E associe un élément x de X).

Nombres cardinaux

A la base de toute tentative de dénombrement se trouve la notion d'**équipotence :** deux ensembles A et B sont équipotents s'il existe une bijection de A sur B.

Cette relation est manifestement réflexive, symétrique et transitive, mais nous ne pouvons pas parler de relation d'équivalence, car nous n'avons défini les relations binaires que sur un ensemble, et l'ensemble de tous les ensembles n'existe pas.

Par exemple, l'ensemble des nombres pairs est équipotent à l'ensemble des nombres impairs (considérer l'application qui à $2n$ associe $2n-1$ pour $n \geqslant 1$) ; l'ensemble des nombres pairs est équipotent à $\mathbb{N}$ (à n on associe $2n$). Ces trois ensembles sont donc équipotents entre eux.

Le second exemple met en évidence une propriété importante de $\mathbb{N}$: l'existence d'une bijection de $\mathbb{N}$ sur une de ses parties propres ($\neq \mathbb{N}$). Pour Dedekind, un **ensemble infini** est un ensemble équipotent à une de ses parties propres. Une autre définition est la suivante : un ensemble E est infini s'il n'existe aucun entier n tel que E soit équipotent à $\{1, 2, ... n\}$. Si on admet l'axiome du choix, les deux définitions sont équivalentes.

De deux ensembles équipotents on dit qu'ils ont même **puissance** ou encore même **cardinal.** On notera $\overline{\overline{A}}$ le cardinal de A (notation de Cantor).

Si A est un ensemble fini, par définition il existe un entier n tel que A soit équipotent à $\{1, 2, ..., n\}$. Si B est un ensemble fini, il est équipotent à $\{1, 2, ..., n'\}$. Alors A est équipotent à B si, et seulement si $n = n'$. On montre que l'on peut identifier $\overline{\overline{A}} = \overline{\overline{B}}$ avec n.

Si A n'est pas un ensemble fini, on dit que son cardinal est **transfini.**

Opérations sur les cardinaux

Soit a et b deux cardinaux, A et B deux ensembles tels que $\overline{\overline{A}} = a$, $\overline{\overline{B}} = b$, $A \cap B = \varnothing$ (cette dernière hypothèse n'est nécessaire que pour la définition de la somme), on pose :

$a + b = \overline{\overline{A \cup B}}$

$a \cdot b = \overline{\overline{A \times B}}$

$a^b = \overline{\overline{\mathcal{F}(B, A)}}$; $\mathcal{F}$ (B, A) est l'ensemble des applications de B dans A.

Dans le cas où a et b sont des nombres finis, on retrouve les opérations usuelles, l'intérêt d'une telle méthode est d'étendre ces opérations aux nombres transfinis.

Pour tous les *nombres cardinaux, finis* ou *non,* les règles de calcul suivantes sont encore valides :

$(a + b) + c = a + (b + c)$
$(a \cdot b) \cdot c = a \cdot (b \cdot c)$
$a + b = b + a$
$a \cdot b = b \cdot a$
$a \cdot (b + c) = a \cdot b + a \cdot c$
$a^c \cdot b^c = (a \cdot b)^c$
$a^b \cdot a^c = a^{b+c}$
$(a^b)^c = a^{b \cdot c}$

Vérifions la dernière égalité. Soit A, B, C des ensembles tels que $\overline{\overline{A}} = a$, $\overline{\overline{B}} = b$, $\overline{\overline{C}} = c$; l'égalité sera vraie si et seulement si $\mathcal{F}(C, \mathcal{F}(B, A))$ est équipotent à $\mathcal{F}(B \times C, A)$. Considérons un élément générique du premier ensemble : à un élément z de C, on associe une application qui, à l'élément y de B, fait correspondre x dans A. La connaissance de z et y détermine donc x de façon unique, ce qui définit une application de $B \times C$ dans A. Il y a donc correspondance biunivoque entre $\mathcal{F}(C, \mathcal{F}(B, A))$ et $\mathcal{F}(B \times C, A)$.

Comparaison de cardinaux

Soit a et b deux cardinaux ; on dira que a est inférieur ou égal à b ($a \leqslant b$) s'il existe une injection d'un ensemble de cardinal a dans un ensemble de cardinal b.

On démontre le résultat fondamental par le théorème suivant.

SECONDE PARTIE DE LA DISME DE L'OPE- RATION.

PROPOSITION I, DE L'ADDITION.

Estant donnez nombres de Disme à ajouster : Trouver leur somme :

Explication du donné. Il y a trois ordres de nombres de Disme, desquels le premier 27 ⓪ 8 ① 4 ② 7 ③, le deuxiesme 37 ⓪ 8 ① 7 ② 5 ③, le troisiesme 875 ⓪ 7 ① 8 ② 2 ③.

Explication du requis. Il nous faut trouver leur somme. *Construction.* On mettra les nombres donnez en ordre comme ci joignant, les aioustant selon la vulgaire maniere d'aiouster nombres entiers, en ceste sorte :

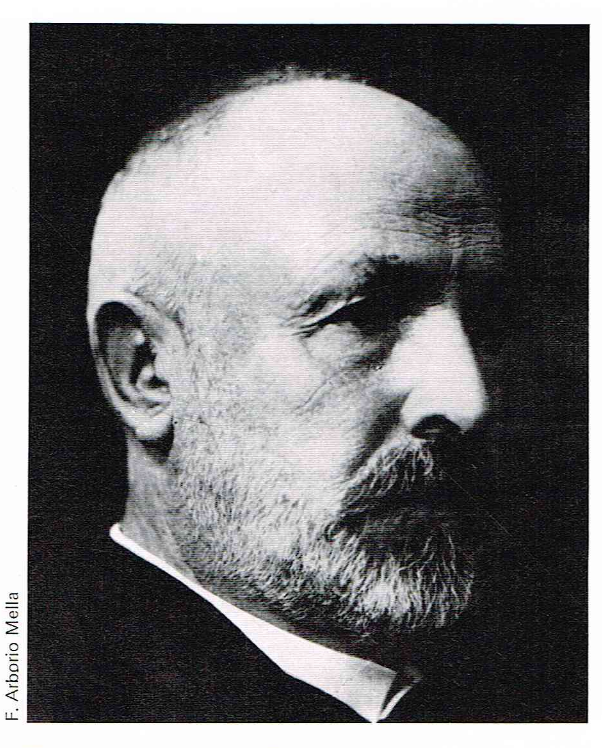
	⓪	①	②	③
2	7	8	4	7
3	7	6	7	5
8 7 5	7	8	2	
9 4 1	3	0	4	

Donne somme (par le 1 probleme de l'Arithmetique) 941304, qui sont (ce que demonstrent les signes dessus les nombres) 941 ⓪ 3 ① 0 ② 4 ③. Ie di, que les mesmes sont la somme requise. *Demonstration.* Les 27 ⓪ 8 ① 4 ② 7 ③ donnez, font (par la 3e definition) 27 $\frac{8}{10}$, $\frac{4}{100}$, $\frac{7}{1000}$, ensemble 27 $\frac{847}{1000}$, & par mesme raison les 37 ⓪ 6 ① 7 ② 5 ③ vallent 37 $\frac{675}{1000}$, & les 875 ⓪ 7 ① 8 ② 4 ③ feront 875 $\frac{782}{1000}$, lesquels trois nombres, comme 27 $\frac{847}{1000}$, 37 $\frac{675}{1000}$, 875 $\frac{782}{1000}$, font ensemble (par le 10e probleme de l'Arith.) 941 $\frac{304}{1000}$, mais autant vaut aussi la somme 941 ⓪ 3 ① 0 ② 4 ③, c'est

Palais de la Découverte, Paris

▲ *Page extraite de l'œuvre de Stevin (édition 1634),* **l'Introduction des fractions décimales,** *dont la première édition a été publiée à Leyde (Hollande) en 1585.*

▶ *Le mathématicien russe* Georg Cantor *(1845-1918) ; il est surtout connu comme créateur de la* théorie des ensembles.



F. Arborio Mella

Théorème de Cantor-Bernstein

Soit A et B deux ensembles. S'il existe une injection de A dans B et une injection de B dans A, alors il existe une bijection de A sur B.

Il en résulte que la relation que nous venons de définir entre cardinaux est antisymétrique. Elle est aussi réflexive et transitive, mais ce n'est pas une relation d'ordre, car, nous le verrons plus loin, il n'y a pas d'ensemble de tous les cardinaux.

La puissance du dénombrable

Le cardinal de l'ensemble $\mathbb{N}$ sera noté comme l'a fait Cantor par le symbole $\aleph_0$ (aleph zéro).

Un ensemble équipotent à $\mathbb{N}$ est dit **dénombrable.** Pour mettre en évidence la bijection entre un ensemble A et $\mathbb{N}$, il suffit de réussir à numéroter les éléments de A. Le schéma suivant indique un procédé permettant cette numérotation pour un produit de deux ensembles dénombrables :

$$\begin{array}{ccccccccc}
a_{11} & \rightarrow & a_{12} & & a_{13} & \rightarrow & a_{14} & & \dots \\
 & \swarrow & & \nearrow & & \swarrow & & \nearrow & \\
a_{21} & & a_{22} & & a_{23} & & a_{24} & & \dots \\
\downarrow & \nearrow & & \swarrow & & \nearrow & & \swarrow & \\
a_{31} & & a_{32} & & a_{33} & & a_{34} & & \dots \\
 & \swarrow & & \nearrow & & \swarrow & & \nearrow & \\
a_{41} & & a_{42} & & a_{43} & & a_{44} & & \dots \\
\downarrow & \nearrow & & \swarrow & & \nearrow & & & \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots
\end{array}$$

On démontre ainsi que tout produit cartésien fini d'ensembles finis ou dénombrables est dénombrable, on a les règles de calcul suivantes :

$\aleph_0 + \aleph_0 = \aleph_0$; $n\aleph_0 = \aleph_0$ n entier naturel
$\aleph_0 \cdot \aleph_0 = \aleph_0$; $\aleph_0^n = \aleph_0$ n entier naturel

On verra par la suite (voir *Nombres*) de nombreux exemples d'ensembles dénombrables.

La puissance du continu

Nous allons démontrer que l'ensemble des nombres réels $\mathbb{R}$ n'est pas dénombrable. Le cardinal de $\mathbb{R}$ est noté **C (puissance du continu).** Comme $\mathbb{R}$ contient $\mathbb{N}$, on a nécessairement $\mathbf{C} > \aleph_0$.

Tout d'abord, remarquons que, pour tout ensemble A, $\overline{\overline{\mathfrak{P}(A)}} = 2^{\overline{\overline{A}}}$. Ceci résulte de la bijection que l'on peut établir entre $\mathfrak{P}(A) = \mathcal{F}(A, \{0, 1\})$; il suffit d'associer à toute partie X de A l'application définie sur A qui, à un élément y, associe 1 si y appartient à A, 0 sinon.

Il existe une injection évidente de A dans $\mathfrak{P}(A)$, celle qui, à l'élément a de A, associe $\{a\}$; nous allons démontrer (par le théorème suivant) qu'il n'existe pas de surjection de A sur $\mathfrak{P}(A)$, il en résultera que $\overline{\overline{A}} < 2^{\overline{\overline{A}}}$.

Théorème de Cantor

Si A est un ensemble, l'ensemble $\mathfrak{P}(A)$ des parties de A n'est pas équipotent à A.

Soit f une application de A dans $\mathfrak{P}(A)$ et soit $X = \{a \in A \mid a \notin f(a)\}$.

Pour démontrer que f n'est pas surjective, il suffit de prouver que $X \notin \mathrm{Im} f$. Faisons un raisonnement par l'absurde. Supposons $f^{-1}(\{X\}) \neq \varnothing$, alors il existe $x \in A$ tel que $f(x) = X$.

Deux cas sont à envisager : ou bien $x \notin X$, il s'ensuit que $x \notin f(x)$, et par définition de X, $x \in X$, ou bien $x \in X$, donc $x \notin f(x)$, et comme $f(x) = X$, $x \notin X$. Les deux cas mènent à une contradiction ; donc $f^{-1}(\{X\}) = \varnothing$ et f n'est pas surjective.

En remarquant que $\overline{\overline{\mathcal{F}(\mathbb{N}, \{0, 1\})}} = 2^{\aleph_0}$, et en se servant du développement binaire des nombres réels (voir *Nombres*) pour établir la bijection entre $\mathbb{R}$ et $\mathcal{F}(\mathbb{N}, \{0, 1\})$, ensemble des suites composées uniquement de 0 et de 1, on démontre que $\mathbf{C} = 2^{\aleph_0}$ d'où l'inégalité stricte :

$$\aleph_0 < \mathbf{C} = 2^{\aleph_0}$$

ce qui prouve que $\mathbb{R}$ n'est pas dénombrable.

Indiquons les règles de calcul concernant $\aleph_0$ et $\mathbf{C}$:

$\mathbf{C} + \mathbf{C} = \mathbf{C}$; $n\mathbf{C} = \mathbf{C}$ $n \in \mathbb{N}$
$\mathbf{C} \cdot \mathbf{C} = \mathbf{C}$; $\mathbf{C}^n = \mathbf{C}$
$\aleph_0 \cdot \mathbf{C} = \mathbf{C} \cdot \aleph_0 = \mathbf{C}$
$n^{\aleph_0} = \mathbf{C}$, $\aleph_0^{\aleph_0} = \mathbf{C}$, $\mathbf{C}^{\aleph_0} = \mathbf{C}$ $(2 \leqslant n < \aleph_0)$
$n^{\mathbf{C}} = 2^{\mathbf{C}}$, $\aleph_0^{\mathbf{C}} = 2^{\mathbf{C}}$, $\mathbf{C}^{\mathbf{C}} = 2^{\mathbf{C}}$, $(2^{\mathbf{C}})^{\mathbf{C}} = 2^{\mathbf{C}}$.

Nous nous limiterons à quelques démonstrations :

$$\mathbf{C}^{\aleph_0} = (2^{\aleph_0})^{\aleph_0} = 2^{\aleph_0 \cdot \aleph_0} = 2^{\aleph_0} = \mathbf{C}$$

de même $\mathbf{C}^{\mathbf{C}} = (2^{\aleph_0})^{\mathbf{C}} = 2^{\aleph_0 \cdot \mathbf{C}} = 2^{\mathbf{C}}$

La relation $\mathbf{C}^n = \mathbf{C}$ exprime le fait que $\mathbb{R}^n$ peut être mis en bijection avec $\mathbb{R}$. Par ailleurs, les relations $\mathbf{C} < 2^{\mathbf{C}}$ et $2^{\mathbf{C}} = \mathbf{C}^{\mathbf{C}}$ montrent que l'ensemble des fonctions réelles à valeurs réelles a une puissance strictement supérieure à celle de $\mathbb{R}$.

La relation fondamentale $\overline{\overline{A}} < 2^{\overline{\overline{A}}}$ met en évidence l'existence d'une suite strictement croissante de cardinaux transfinis qui eux-mêmes succèdent aux cardinaux finis. Une telle suite s'écrit :

(a) $\aleph_0,\quad 2^{\aleph_0},\quad 2^{2^{\aleph_0}},\quad 2^{2^{2^{\aleph_0}}},\quad \dots$

Nous pouvons alors considérer le nombre cardinal Ω_0 associé à la réunion d'une famille d'ensembles de cardinaux $\aleph_0$, $2^{\aleph_0}$, $2^{2^{\aleph_0}}$, ... ; Ω_0 est supérieur à tous ces cardinaux, et à la suite (a) on peut faire succéder la suite :

(b) $\Omega_0,\quad 2^{\Omega_0},\quad 2^{2^{\Omega_0}},\quad 2^{2^{2^{\Omega_0}}},\quad \dots$

On construit ainsi une suite transfinie, strictement croissante, de cardinaux (c). On démontre qu'une telle suite n'est pas un ensemble :

(c) $\aleph_0, 2^{\aleph_0}, 2^{2^{\aleph_0}}, \dots, \Omega_0, 2\Omega_0, 2^{2^{\Omega_0}}, \dots, \Omega_1, 2\Omega_1, 2^{2^{\Omega_1}}, \dots$

L'axiome du choix permet de démontrer que tout cardinal transfini, ou coïncide avec un élément de (c) ou peut être inséré entre deux éléments de (c).

L'hypothèse du continu, conjecturé par Cantor lui-même, consiste à affirmer qu'il n'y a aucun nombre cardinal entre $\aleph_0 = 2^{\aleph_0}$.

L'hypothèse du continu généralisé assure qu'il n'y a aucun nombre cardinal entre a et 2^a.

Une autre approche de ce problème va être exposée avec l'étude des nombres ordinaux.

Nombres ordinaux

Pour des ensembles totalement ordonnés, la notion d'ensembles équipotents peut être renforcée par celle d'ensembles isomorphes. Deux ensembles totalement ordonnés $(M, \leqslant)$ et $(M_*, \leqslant_*)$ sont isomorphes (ou semblables) s'il existe une bijection f de M sur M_* telle que, pour tout couple (x, y) d'éléments de M $x \leqslant y$ si, et seulement si, $f(x) \leqslant_* f(y)$.

La relation ainsi définie est réflexive, symétrique et transitive, mais, pas plus que pour la relation d'équipotence, nous n'avons ici une relation d'équivalence. Il faudrait pour cela que l'ensemble de tous les ordinaux existât, ce qui est faux.

De deux ensembles isomorphes on dit qu'ils ont même **type d'ordre**. On note $\overline{(M, \leqslant)}$ le type d'ordre de l'ensemble M ordonné par la relation $\leqslant$.

Les **nombres ordinaux** seront les types d'ordres des ensembles bien ordonnés. Un ensemble ordonné est par définition **bien ordonné** si toute partie non vide de cet ensemble admet un plus petit élément.

L'ensemble $\mathbb{N}$ muni de la relation d'ordre usuelle est un ensemble bien ordonné. On désigne par ω le type d'ordre correspondant.

L'ensemble $\mathbb{N}$ muni des *ordres* définis par :

(1) $1 \leqslant 2 \leqslant 3 \ldots\ldots\ldots\ldots \leqslant 0$
(2) $2 \leqslant 3 \leqslant 4 \ldots\ldots\ldots\ldots \leqslant 0 \leqslant 1$
(3) $0 \leqslant 2 \leqslant 4 \leqslant \ldots\ldots \leqslant 1 \leqslant 3 \leqslant 5 \ldots$
(4) $0 \leqslant 3 \leqslant 6 \leqslant \ldots\ldots$
$\leqslant 1 \leqslant 4 \leqslant 7 \leqslant \ldots \leqslant 2 \leqslant 5 \leqslant 8 \ldots$

est encore bien ordonné, mais les nombres ordinaux correspondants sont différents.

(1) possède un plus grand élément alors que $(\mathbb{N}, \leqslant)$ n'en a pas.

$\mathbb{N}$ peut être bien ordonné d'une infinité de façons créant une infinité de nombres ordinaux différents; cette propriété s'étend à tout ensemble de puissance supérieure à celle de $\mathbb{N}$.

Notons qu'un ensemble peut être totalement ordonné sans être bien ordonné (la réciproque est fausse). Il suffit de considérer $\mathbb{N}$ ordonné par :

$$0 \leqslant 2 \leqslant 4 \leqslant 6 \ldots \leqslant 7 \leqslant 5 \leqslant 3 \leqslant 1$$

Le sous-ensemble des nombres pairs n'a pas de plus petit élément.

Mais cela ne peut se produire que pour des ensembles infinis. Deux ensembles finis de même cardinal n et totalement ordonnés sont semblables. On identifiera leur nombre ordinal à n, indiquant ainsi que le résultat du dénombrement dans le cas fini est indépendant de l'ordre choisi pour dénombrer. Il n'en est pas de même pour les ensembles infinis.

Opérations sur les ordinaux

Nous nous limiterons à définir la somme de deux **ordinaux.**

Étant donné deux ordinaux α et β, on considère les ensembles bien ordonnés $(A, \leqslant_A)$, $(B, \leqslant_B)$ tels que :

$$\overline{(A, \leqslant_A)} = \alpha \quad \text{et} \quad \overline{(B, \leqslant_B)} = \beta, \quad A \cap B = \varnothing.$$

Alors $\alpha + \beta = \overline{(A \cup B, \leqslant_{AB})}$ où $\leqslant_{AB}$ est une relation de bon ordre telle que $x \leqslant_{AB} y$ si et seulement si $x \leqslant_A y$ ou $x \leqslant_B y$ au lieu de $x \in A$ et $y \in B$. Tous les éléments de A précèdent ceux de B.

On démontre que $\alpha + \beta$ est indépendant des ensembles A et B choisis.

L'addition des nombres ordinaux finis a les mêmes propriétés que l'addition des nombres finis. L'addition de nombres ordinaux n'est pas, en général, commutative.

Considérons les ensembles $\mathbb{N}$ et $\mathbb{N}^* = \mathbb{N} - \{0\}$ munis de la relation d'ordre usuelle. Ils sont semblables. $(\mathbb{N}, \leqslant)$ est la réunion ordonnée des ensembles bien ordonnés $(\{0\}, \leqslant)$ et $(\mathbb{N}^*, \leqslant)$. Le nombre ordinal de $\{0\}$ est 1, celui de $(\mathbb{N}, \leqslant)$ est ω; on a donc : $1 + \omega = \omega$.

La réunion ordonnée des ensembles bien ordonnés $(\mathbb{N}^*, \leqslant)$ et $(\{0\}, \leqslant)$ est l'ensemble ordonné (1) ci-dessus. Son type d'ordre $\omega + 1$ est différent de ω.

D'où $1 + \omega \neq \omega + 1$.

Les nombres ordinaux des ensembles (2), (3), (4) sont respectivement

$$\omega + 1, \quad \omega + 2, \quad \omega + \omega + \omega.$$

Comparaison d'ordinaux

Étant donné deux ensembles bien ordonnés $(A, \leqslant_A)$ et $(B, \leqslant_B)$, le second est un **segment initial** du premier si B est un sous-ensemble de A qui, chaque fois qu'il contient un élément de A, contient tous ses prédécesseurs, l'ordre $\leqslant_B$ étant la restriction de $\leqslant_A$ à l'ensemble B.

α et β étant deux ordinaux, on dira que β est inférieur à α, $(\beta \leqslant_0 \alpha)$, s'il existe des ensembles bien ordonnés $(A, \leqslant_A)$ et $(B, \leqslant_B)$ tels que $\alpha = \overline{(A, \leqslant_A)}$ et $\beta = \overline{(B, \leqslant_B)}$ avec $(B, \leqslant_B)$ segment initial de $(A, \leqslant_A)$.

Cette relation est réflexive, antisymétrique et transitive, mais, comme les nombres ordinaux ne forment pas un ensemble, on ne peut pas parler de relation d'ordre. Cependant on démontre que, sur tout ensemble de nombres ordinaux, la relation $\leqslant_0$ induit un bon ordre, donc un ordre total.

Un ensemble K d'ordinaux qui, lorsqu'il contient un ordinal, contient tous ses prédécesseurs pour la relation $\leqslant_0$, est appelé segment initial d'ordinaux, on le note $(K, \leqslant_0)$. L'ensemble $P(\alpha)$ des ordinaux qui précèdent l'ordinal α pour $\leqslant_0$ est un segment initial d'ordinaux, et réciproquement tout segment initial d'ordinaux peut se mettre sous la forme $P(\alpha)$. La correspondance qui à α associe $P(\alpha)$ est donc une correspondance biunivoque entre ordinaux et segments initiaux d'ordinaux, et on démontre qu'en général le type d'ordre de $P(\alpha)$ est α. On a ainsi :

$$\overline{P(\omega)} = \omega$$

(il suffit de vérifier que ω est précédé par et seulement par tous les ordinaux 0, 1, 2, 3, ...)

$$\overline{P(\omega + 1)} = \omega + 1$$

($\omega + 1$ est précédé par et seulement par 0, 1, 2, ..., ω)

$$\overline{P(n)} = n$$
$$\overline{P(0)} = 0$$

Étant donné un segment initial d'ordinaux, il existe un nombre ordinal qui domine tous les ordinaux du segment, c'est le type d'ordre du segment lui-même. Ajoutant un tel nombre ordinal au segment, on obtient un nouveau segment initial d'ordinaux qui contient strictement le premier. On réitère le procédé. On est ainsi amené à construire une suite transfinie strictement croissante de nombres ordinaux, analogue en quelque sorte à la suite transfinie des cardinaux du paragraphe précédent.

On peut classer les nombres ordinaux en deux catégories : la première constituée par les ordinaux qui ont un antécédent immédiat dans la suite transfinie des ordinaux, la seconde constituée par les ordinaux qui n'en ont pas. A la première catégorie appartiennent les ordinaux finis supérieurs à 0 et $\omega + 1$, $\omega + 2$...; à la seconde appartiennent ω, $\omega + \omega = \omega_2$, ... On peut représenter la suite transfinie des idéaux de la façon suivante :

0	1	2	3 . . . . ω
	$\omega + 1$	$\omega + 2$	$\omega + 3$ . . . . . ω_2
	$\omega_2 + 1$	$\omega_2 + 2$	$\omega_2 + 3$ . . . . . ω_3
	.	.	. .
	.	.	. .
	.	.	. .
	$\omega_n + 1$	$\omega_n + 2$	ω_n
	. . . .	. . . .	. . . .

A tout nombre ordinal est associé un cardinal bien défini, celui de tout ensemble ayant un type d'ordre égal au nombre ordinal. Les ensembles qui ont pour ordinaux ω, $\omega + 1$, $\omega + 2$, ..., ω_2 ont tous la puissance du dénombrable. A tout nombre cardinal donné correspond toute une classe de nombres ordinaux. Nous n'approfondirons pas plus ici ces propriétés.

Signalons cependant que l'on peut démontrer l'équivalence entre l'axiome du choix et l'axiome de Zermelo : « Tout ensemble peut être bien ordonné ».

Nous terminerons par une citation de Cantor au sujet des relations $\omega + 1 = \omega + 1$ et $1 + \omega = \omega$: « Nous l'avons vu, c'est la position relative du fini par rapport à l'infini qui importe; si le fini précède il est absorbé par l'infini et disparaît, s'il suit l'infini il subsiste et transforme celui-ci en un nouvel infini différent du premier. »

BIBLIOGRAPHIE

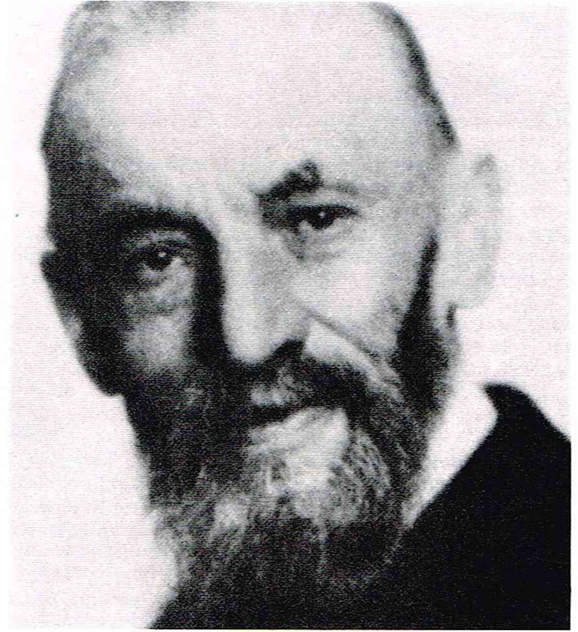
BOURBAKI, *Théorie des ensembles,* Paris, Hermann, 1970; *Éléments d'histoire des mathématiques,* Paris, Hermann, 1960. - BOUVIER, *la Théorie des ensembles,* coll. « Que sais-je ? », Paris, P.U.F., 1969. - BREUER, *Initiation à la théorie des ensembles,* Paris, Dunod, 1969. - CAVAILLES, *Philosophie mathématique,* Paris, Hermann, 1962. - DESANTI, *la Philosophie silencieuse,* Paris, Seuil, 1975. - EXBRAYAT et MAZET, *Algèbre 1,* Paris, Hatier, 1971. - GODEMENT, *Cours d'algèbre,* Paris, Hermann, 1966. - HALMOS, *Introduction à la théorie des ensembles,* Paris, Gauthier-Villars, 1967. - LE LIONNAIS, *les Grands Courants de la pensée mathématique,* Paris, Blanchard, 1962.

▶ *Sur cette planche, tirée de l'*Histoire des mathématiques* de Montucla, divers systèmes de numération et caractères arithmétiques utilisés par un certain nombre d'auteurs à différentes époques. On remarquera que chez Boèce il manque encore une notation pour le zéro.*

Anciens Caracteres Arithmetiques

1. *Notes de Boece.*
2. *De Planude.*
3. *Caracteres d'Alsephadi.*
4. *Chiffres de Sacro Bosco.*
5. *De Roger Bacon.*
6. *Des Indiens Modernes.*
7. *Chiffres Modernes.* 1 2 3 4 5 6 7 8 9 10
8. *Nombre d'Alsephadi.*

Ciccione



Buscaglia

▶ *Le logicien et mathématicien italien,* Guiseppe Peano *(1858-1932); on lui doit notamment un exposé axiomatique de la théorie des nombres entiers.*

LES NOMBRES

Les entiers naturels

La notion de nombre entier qui nous est si familière n'a été élaborée que très lentement au cours de l'histoire. Compter, c'est-à-dire mettre en correspondance les objets de deux collections, est une opération qui ne nécessite pas l'usage de mots spécifiques. Lévy-Bruhl a remarqué que dans certaines sociétés primitives ces mots sont tout à fait ignorés sans pour autant rendre impossible la démarche qu'ils impliquent.

Les mots numérateurs ne suffisent pas non plus à établir l'existence de nombres abstraits, ils ont été longtemps marqués par des valeurs magiques ou affectives dont il reste encore des traces de nos jours (le nombre 13). Avec Pythagore, les nombres se détachent de représentations réelles pour être pensés comme des arrangements privilégiés de points. C'est seulement au XIXe siècle, avec l'essor de l'axiomatique, que les mathématiciens s'affranchissent de toute représentation pour définir les nombres dits naturels.

On doit à Peano (1858-1932) et à R. Dedekind (1831-1916) un exposé axiomatique de la théorie des nombres entiers. Nous allons montrer comment on peut déduire l'arithmétique élémentaire de cinq axiomes (qui sont, à quelques modifications secondaires près, ceux qui ont été formulés par Peano en 1889) :

I) 1 est un nombre naturel;
II) pour tout nombre naturel n, il existe un nombre naturel n' unique, appelé *successeur* de n;
III) 1 n'est le successeur d'aucun nombre;
IV) deux nombres naturels différents ont deux successeurs différents;
V) une propriété vraie pour le nombre 1, et qui, si elle est vraie pour un nombre naturel, est alors vraie pour son successeur, est vraie pour tout entier naturel.

L'axiome V est très important, car il fonde le raisonnement par récurrence. Ainsi, il va servir à définir par récurrence l'*addition :*

Le système

(1) $$\begin{cases} m+1=m' & \text{(successeur de } m\text{)} \\ m+n'=(m+n)' & \text{(successeur de } m+n\text{)} \end{cases}$$

permet de définir de façon unique une opération, $+$, qui, à tout couple (m, n) d'entiers naturels, associe un entier naturel unique $m+n$, appelé somme de m et de n.

Tout d'abord, remarquons grâce à l'axiome **IV** que $(m+n)'$ est unique, donc également $m+n'$. Soit alors P (n) la propriété : « $m+n$ est défini de façon unique par le système **(1)** pour tout nombre naturel n »; P (1) est vraie. En outre, d'après la remarque précédente, si P (n) est vraie, alors P (n') l'est aussi. L'axiome **V** permet d'affirmer que P (n) est vraie pour tout entier naturel n, c'est-à-dire que $m+n$ est défini de façon unique pour tout couple m et n.

De même, on définira par récurrence la *multiplication :* le système :

(2) $$\begin{cases} m \cdot 1 = m \\ m \cdot n' = (m \cdot n) + m \end{cases}$$

permet de définir de façon unique une opération, $\cdot$, qui à tout couple (m, n) d'entiers naturels associe un entier naturel unique $m \cdot n$ (noté aussi mn) appelé *produit* de m et de n. La démonstration est analogue à la précédente.

Toujours en utilisant l'axiome **V**, on démontre que l'addition et la multiplication des entiers naturels ont les propriétés suivantes :

(Aa) $(a+b)+c=a+(b+c)$
(associativité de l'addition)

(Ca) $a+b=b+a$
(commutativité de l'addition)

(Am) $(a \cdot b) \cdot c = a \cdot (b \cdot c)$
(associativité de la multiplication)

(Cm) $a \cdot b = b \cdot a$
(commutativité de la multiplication)

(D_1) $a \cdot (b+c) = a \cdot b + a \cdot c$
(distributivité de la multiplication par rapport à l'addition)

(D_2) $(a+b) \cdot c = a \cdot c + b \cdot c$
(distributivité de la multiplication par rapport à l'addition)



◀ *L'*arithmétique *dans une miniature du* De arithmetica arte *qui fait partie du traité de Cassiodore sur les arts libéraux (Paris, Bibliothèque nationale).*

Giraudon

Remarque

Les axiomes de Peano tels que nous les avons énoncés définissent l'ensemble $\mathbb{N}^* = \{1, 2, 3, ...\}$; une autre version fait intervenir le zéro, il suffit de le substituer à 1 dans les axiomes (et alors $0' = 1$). Dans le système **(1)** on remplace la première égalité par $m + 0 = m$, et dans le système **(2)** par $m \cdot 0 = 0$. Toutes les autres considérations du paragraphe restent inchangées.

Ordre sur $\mathbb{N}$

L'ensemble $\mathbb{N} = \{0, 1, 2, 3, ...\}$ des entiers naturels est ordonné par la relation suivante :

$a \leqslant b$ (a est inférieur ou égal à b) ou encore $b \geqslant a$ (b est supérieur ou égal à a) si, et seulement si il existe un nombre naturel c tel que $a + c = b$; cette relation est manifestement réflexive, antisymétrique et transitive. C'est une *relation d'ordre total,* car, pour tout couple (a, b), l'une ou l'autre des deux assertions $a \leqslant b$ ou $b \geqslant a$ est vraie.

Cet ordre est compatible avec l'addition et la multiplication, ce qui revient à dire :

si $a \leqslant b$ alors $a + c \leqslant b + c$ et $ac \leqslant bc$; en effet $a \leqslant b$ équivaut à $b = a + d$, alors $b + c = a + d + c = a + c + d$ (d'après Aa et Ca), donc $a + c \leqslant b + c$; de même, $bc = (a + d)\ c = ac + dc$ et $ac \leqslant bc$.

On démontre que l'équation $a + x = b$, pour deux entiers naturels quelconques a et b, admet au plus une solution dans $\mathbb{N}$. Cette solution existe si, et seulement si $a \leqslant b$, on la note $b - a$, c'est la différence entre b et a' et l'opération correspondante est la *soustraction.*

Les entiers relatifs

Nous venons de remarquer que la soustraction ne peut pas toujours être effectuée pour deux entiers quelconques. Pour combler cette lacune, nous allons construire un nouvel ensemble, noté $\mathbb{Z}$, celui des entiers relatifs.

Cette extension ne peut se faire sans précaution. Les opérations que l'on définira sur le nouvel ensemble doivent coïncider avec celles qui sont déjà connues sur $\mathbb{N}$, considéré comme partie de $\mathbb{Z}$. Nous allons montrer de façon rigoureuse comment construire $\mathbb{Z}$ à partir des entiers naturels.

Le problème que nous nous proposons de résoudre est celui de donner un sens à $a - b$ quels que soient les entiers a et b et pas seulement dans le cas où $a \geqslant b$. Nous savons déjà que la proposition suivante :

(1) $a - b = c - d$ si, et seulement si $a + d = b + c$ est vraie pour $a \geqslant b$ et $c \geqslant d$, c'est-à-dire quand $a - b$ et $c - d$ sont des entiers naturels. C'est cette restriction qu'il nous faut abolir.

Considérons la relation suivante sur $\mathbb{N} \times \mathbb{N}$:

$(a, b) \sim (c, d)$ si, et seulement si $a + d = b + c$ où la soustraction n'intervient plus.

Cette relation est une relation d'équivalence, c'est-à-dire une relation

— *réflexive :* $(a, b) \sim (a, b)$

— *symétrique:* si $(a, b) \sim (b, a)$ alors $(b, a) \sim (a, b)$

— *transitive :* si $(a, b) \sim (a', b')$ et $(a', b') \sim (a'', b'')$ alors $(a, b) \sim (a'', b'')$.

La vérification est immédiate.

Nous pouvons donc obtenir une partition de l'ensemble $\mathbb{N} \times \mathbb{N}$, ensemble des couples d'entiers naturels, en *classes d'équivalence* deux à deux disjointes.

Ainsi : $(2, 3) \sim (7, 8)$ car $2 + 8 = 7 + 3$

$(2, 3)$ et $(7, 8)$ appartiennent à la même classe notée indifféremment $[2, 3]$ ou $[7, 8]$.

On note $[a, b]$ la classe à laquelle appartient le couple (a, b) que l'on appellera *représentant* de la classe.

Nous désignerons alors par $\mathbb{Z}$ l'ensemble des classes d'équivalence :

$$\mathbb{Z} = \{[a, b] \mid a, b \in \mathbb{N}\}$$

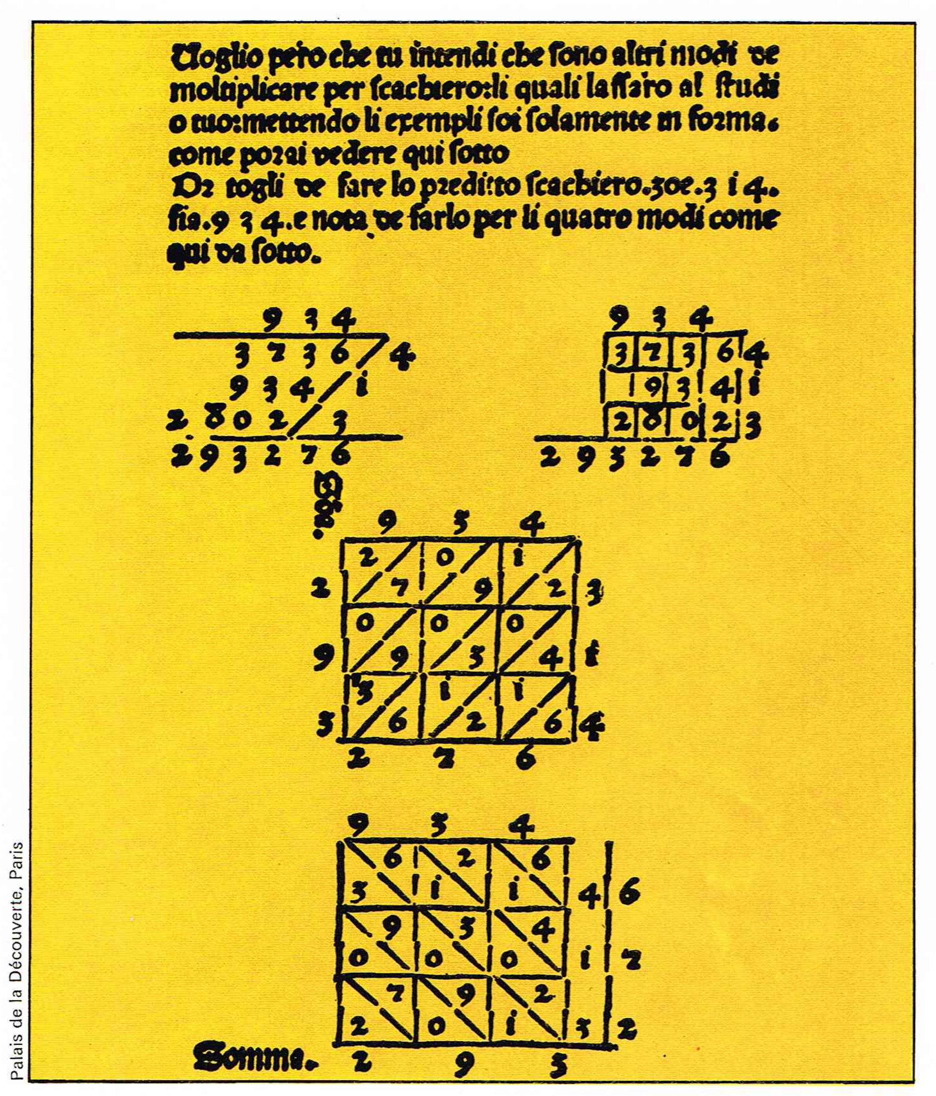
Il reste à définir dans $\mathbb{Z}$ les opérations, somme, produit, qui l'ont été dans $\mathbb{N}$. Pour cela, on pose

(2) $[a, b] + [c, d] = [a + c, b + d]$

(3) $[a, b] \cdot [c, d] = [ac + bd, ad + bc]$.

Par convention, la multiplication l'emporte sur l'addition, ce qui évite l'usage excessif de parenthèses dans une expression du type $ac + bd$, par exemple.

Les définitions **(2)** et **(3)** ne sont acceptables que si elles sont indépendantes des représentants cités ; en d'autres termes, si $(a', b') \sim (a, b)$ et $(c', d') \sim (c, d)$ implique


Voglio pero che tu intendi che sono altri modi de moltiplicare per ſcachiero: li quali laſſaro al ſtudio tuo: mettendo li exempli ſoi ſolamente in forma. come porai vedere qui ſotto
Or togli de fare lo predito ſcachiero. zoe. 3 1 4. fia. 9 3 4. e nota de farlo per li quatro modi come qui va ſotto.

Somma.

Palais de la Découverte, Paris

▲ ***Page extraite d'un ouvrage d'arithmétique édité à Trévise en 1478 et montrant diverses formes de*** **multiplication.**

$[a', b'] + [c', d'] = [a, b] + [c, d] = [a + c, b + d]$

et

$[a', b'] \cdot [c', d'] = [a, b] \cdot [c, d] = [ac + bd, ad + bc]$.

On dira alors que les opérations considérées sont compatibles avec la relation d'équivalence.

Vérifions la première propriété. Cela revient à montrer que $[a' + c', b' + d'] = [a + c, b + d]$. Par hypothèse, $(a', b') \sim (a, b)$, ce qui équivaut à $a' + b = b' + a$, de même $c' + d = d' + c$, il en résulte

$(a' + b) + (c' + d) = (b' + a) + (d' + c)$, soit encore $(a' + c') + (b + d) = (b' + d') + (c + a)$, d'où le résultat.

Nous nous sommes servi du fait que $[a, b] = [c, d]$ si, et seulement si $(a, b) \sim (c, d)$, c'est-à-dire, $a + d = b + c$; cela va nous permettre de faire la remarque suivante :

Si $b \geqslant a$, l'unique solution de l'équation $a + x = b$ dans $\mathbb{N}$ est parfaitement connue et notée $b - a$; on aura

$a + (b - a) = b = b + 0$ d'où $[a, b] = [0, b - a]$.

Si $a \geqslant b$, alors $a = a + 0 = b + (a - b)$ et $[a, b] = [a - b, 0]$.

Toute classe peut se mettre sous l'une ou l'autre de ces deux formes : $[n, 0]$ ou $[0, n]$.

Si nous considérons la bijection qui, à la classe $[0, n]$, associe l'entier n, nous nous apercevons que c'est un isomorphisme (une bijection qui conserve les structures définies sur chaque ensemble); plus précisément :

$[0, n] + [0, m] = [0, m + n]$

$[0, n] \cdot [0, m] = [0 \cdot m + n \cdot 0, 0 \cdot 0 + n \cdot m] = [0, n \cdot m]$

cela nous permet d'identifier l'ensemble des classes de la forme $[0, n]$ avec l'ensemble $\mathbb{N}$ et de justifier l'égalité $[0, n] = n$ (en toute rigueur, ceci est illégitime car $[0, n]$ appartient à l'ensemble des classes d'équivalence de $\mathbb{N} \times \mathbb{N}$ pour une certaine relation d'équivalence et n appartient à $\mathbb{N}$). On posera $[n, 0] = -n$, et alors

$$\mathbb{Z} = \{n, -n \mid n \in \mathbb{N}\}.$$

$\mathbb{Z}$ muni de l'addition est un groupe commutatif, c'est-à-dire un ensemble muni d'une loi

— *commutative :*

$[a, b] + [c, d] = [c, d] + [a, b]$,

— *associative :*

$([a, b] + [c, d]) + [e, f] = [a, b] + ([c, d] + [e, f])$ ce qui rend inutile l'usage des parenthèses,

— *possédant un élément neutre :* $[0, 0] = 0$

ainsi $[a, b] + [0, 0] = [a + 0, b + 0] = [a, b]$,

— *et telle que chaque élément* $[a, b]$ *possède un symétrique* $[a', b']$ vérifiant $[a, b] + [a', b'] = [0, 0]$; en effet, $[a, b] + [b, a] = [a + b, b + a] = [0, 0]$. En particulier, le symétrique de $[0, n] = n$ sera $[n, 0] = -n$.

Le symétrique de tout élément est unique.

Pour le vérifier, supposons que l'on ait :

$\alpha + (-\alpha) = \alpha + (-\beta) = 0$, alors $(-\alpha) + \alpha + (-\alpha) = (-\alpha) + \alpha + (-\beta)$ d'où $0 + (-\alpha) = 0 + (-\beta)$ et $-\alpha = -\beta$

si $a \leqslant b$ $[a, b] = [0, b - a] = b - a$;

si $a \geqslant b$ $[a, b] = [a - b, 0] = -(a - b)$;

or $[a - b, 0] + [b - a, 0] = [0, 0]$.

Le symétrique de $a - b$ est donc $b - a$, et dans tous les cas on a $[a, b] = b - a$.

Nous sommes maintenant en mesure d'affirmer que l'équation $a + x = b$ admet une solution unique dans $\mathbb{Z}$, quels que soient a et b, à savoir $x = b - a$.

Il nous reste à examiner les propriétés de la multiplication dans $\mathbb{Z}$. Elle est

— *commutative :* $[a, b] \cdot [c, d] = [c, d] \cdot [a, b]$,

— *associative :*

$([a, b] \cdot [c, d]) \cdot [e, f] = [a, b] \cdot ([c, d] \cdot [e, f])$;

en outre

$[a, b] \cdot ([c, d] + [e, f]) = [a, b] \cdot [c, d] + [a, b] \cdot [e, f]$

ce que l'on traduit en disant que la multiplication est distributive par rapport à l'addition. Les démonstrations se font en se reportant aux définitions des opérations sur les classes données plus haut.

$\mathbb{Z}$ muni de ces deux lois est un *anneau commutatif.*

De façon générale *(cf. Structures),* on appelle anneau le triplet $(A, +, \cdot)$ où A est un ensemble sur lequel on a défini deux opérations $+$ et $\cdot$, c'est-à-dire deux applications de $A \times A \to A$ possédant les propriétés suivantes:

— $(A, +)$ est un groupe *abélien* (commutatif) ;

— la multiplication est *associative ;*

— la multiplication est *distributive* à droite et à gauche par rapport à l'addition.

Si la multiplication est commutative, on dit que l'anneau est commutatif. $\mathbb{Z}$ en est un des exemples les plus importants.

On démontre que $\mathbb{Z}$ est le plus petit anneau contenant $\mathbb{N}$ dans lequel l'équation $a + x = b$ admet toujours une solution.

Propriétés des nombres entiers

L'ordre

Parmi d'autres propriétés remarquables, $\mathbb{Z}$ possède celle d'être un ensemble totalement ordonné (étant donné deux éléments quelconques a et b de $\mathbb{Z}$, on a toujours $a \leqslant b$ ou $b \leqslant a$).

Nous allons montrer comment l'ordre total obtenu sur $\mathbb{N}$ induit un ordre total sur $\mathbb{Z}$ respectant la structure d'anneau.

Un anneau A *ordonné* est un anneau sur lequel est définie une structure d'ordre total compatible avec les lois de l'anneau, c'est-à-dire vérifiant :

(01) Deux éléments quelconques a et b sont comparables

$$\forall a \in A, \quad \forall b \in A \quad a \leqslant b \text{ ou } b \leqslant a$$

(02) $\forall c \in A \quad (a \leqslant b \Rightarrow a + c \leqslant b + c)$

(03) $a \leqslant b, \quad c \geqslant 0 \Rightarrow ac \leqslant bc$

Dans un anneau ordonné, on a l'équivalence

(1) $a \leqslant b \Leftrightarrow b - a \geqslant 0.$

Ce qui nous permettra de reformuler la définition d'un anneau ordonné sous une autre forme équivalente qui se révélera utile.

Un anneau ordonné est un anneau pour lequel on peut déterminer un sous-ensemble P, que l'on appellera sous-ensemble des éléments positifs, $(x > 0 \Leftrightarrow x \in P)$, tel que :

(01') pour tout élément a de $A^* = A - \{0\}$ une, et une seule, des relations $a > 0$, $-a > 0$ est vraie ;

(02') $a \geqslant 0, b \geqslant 0 \Rightarrow a + b \geqslant 0$;

(03') $a \geqslant 0, b \geqslant 0 \Rightarrow ab \geqslant 0$.

Supposons A ordonné selon la première définition. Comme l'ordre est total pour tout élément a, $a \geqslant 0$ ou $a \leqslant 0$; si $a \leqslant 0$ alors $a + (-a) \leqslant -a$ d'après **(02)**, donc $0 \leqslant -a$ et **(01')** est vrai. Si $a \geqslant 0$, $b \leqslant a + b$ d'après **(02)**, comme $b \geqslant 0$ la transitivité de la relation d'ordre implique $a + b \geqslant 0$, donc **(02')** est vraie. **(03')** est une conséquence immédiate de **(03)**.

Réciproquement supposons A ordonné selon la seconde définition. On vérifie immédiatement que la relation $a \leqslant b$ si, et seulement si $a - b \geqslant 0$ est une relation d'ordre total. **(02)** se déduit immédiatement de la définition de l'ordre. Enfin, de $a \leqslant b$, $c \geqslant 0$, on déduit $b - a \geqslant 0$, d'où $bc - ac \geqslant 0$ (d'après **03'**) et $ac \leqslant bc$. Il y a bien équivalence entre les deux définitions.

C'est la deuxième définition qui nous servira : $\mathbb{N}$ considéré comme partie de $\mathbb{Z}$ a bien les propriétés requises et permet donc de faire de $\mathbb{Z}$ un anneau ordonné.

La divisibilité

Un des problèmes fondamentaux dans $\mathbb{N}$ comme dans $\mathbb{Z}$ est celui de la résolution de l'équation $ax = b$ et plus généralement celui de l'étude des propriétés de divisibilité dans ces ensembles.

Faisons d'abord quelques remarques :

— il existe dans $\mathbb{Z}$ un élément neutre pour la multiplication, le nombre 1 tel que $a1 = 1a = 1$;

— la multiplication dans $\mathbb{Z}$ est une opération régulière : $ac = bc$, $c \neq 0 \Rightarrow a = b$,

ce qui équivaut à la proposition suivante :

$$ab = 0 \Rightarrow a = 0 \quad \text{ou} \quad b = 0.$$

Comme on le voit en remarquant que $ac = bc$, $c \neq 0$, $\Leftrightarrow (a - b)\, c = 0$, $c \neq 0$. La première propriété exprime que $\mathbb{Z}$ est un anneau unitaire, la deuxième que c'est un anneau intègre.

Si b est un entier quelconque, tout entier a pour lequel l'équation $ax = b$ admet une solution entière est appelé *diviseur de b*. Par la suite, nous emploierons indifféremment les expressions a est un diviseur de b, b est divisible par a, b est un multiple de a, que nous noterons $a \mid b$.

Ainsi, pour tout entier a, $a \mid 0$. Par contre $0 \mid b$ si, et seulement si $b = 0$. L'équation $0x = b$ a une infinité de solutions si $b = 0$, aucune solution si $b \neq 0$. L'équation $ax = b$ a une solution unique si $a \neq 0$ (ceci résulte de la régularité de la multiplication), solution notée $\frac{b}{a}$.

$\frac{b}{a}$ n'est donc définie que pour $a \neq 0$.

Tout entier a possède quatre diviseurs triviaux $1, -1, a, -a$. D'ailleurs, si a est un diviseur de b, c'est-à-dire s'il existe un entier x tel que $b = a \cdot x$, alors $b = (-a)\,(-x)$, ce qui montre que $-a$ est un diviseur de b. Dans la théorie de la division, il n'y a pas lieu de faire une distinction entre deux nombres opposés, et nous ne considérerons que les positifs.

Étudions dans $\mathbb{N}$ la relation $a \mid b$. Elle est réflexive, antisymétrique ($a \mid b$ et $b \mid a \Leftrightarrow a = b$) et transitive ($a \mid b$, $b \mid c \Rightarrow a \mid c$) ; c'est donc une relation d'ordre. On dira que $a \prec b$ (a est antérieur à b) si, et seulement si a divise b. Contrairement à la relation d'ordre usuelle dans $\mathbb{N}$ qui est une relation d'ordre total (a et b sont toujours comparables), cette deuxième relation d'ordre est une relation d'ordre partiel : $2 \prec 12$, car 12 est un multiple de 2 mais 2 et 5 ne sont pas comparables. Étant donné deux entiers a et b, le plus petit entier postérieur à a et b au sens de la relation $\prec$ est appelé *plus petit commun multiple (ppcm)* de a et b ; le plus grand entier antérieur à a et b au sens de la relation $\prec$ est le *plus grand commun diviseur (pgcd)* de a et de b. L'ensemble N ainsi ordonné est un treillis *(cf. Structures)*.

Palais de la Découverte, Paris

▲ ***Une** division par effaçage : page extraite du même ouvrage d'arithmétique (Trévise, 1478).*

La détermination du *pgcd* de deux nombres a et b peut se faire grâce à un procédé dû à Euclide et connu sous le nom d'*algorithme d'Euclide*.

Étant donné deux entiers a et b, il existe toujours un entier q (quotient) et un entier r (reste) tels que :

$a = bq + r \quad 0 \leqslant r < b$ division euclidienne

(il suffit de choisir pour q le plus grand entier tel que $bq \leqslant a$ et de prendre $r = a - bq$)

q et r sont déterminés de façon unique.

Supposons que α soit un diviseur de a et de b ($a \geqslant b$), α divise r, α est donc un diviseur commun de r et de b, le *pgcd* de a et b est donc le même que celui de b et r ; on réitère le procédé jusqu'à ce que l'on trouve un reste nul, l'avant-dernier reste est alors le *pgcd*. Appliquons cet algorithme à la recherche du *pgcd* de 336 et 144 ; on a :

$$336 = 144\,(2) + 48$$
$$144 = 48\,(3) + 0$$

le *pgcd* de 336 et 144 est 48.

On vérifie sans peine que tous les diviseurs communs de 336 et 144 sont exactement ceux de 48.

Deux entiers a et b sont *premiers entre eux* si, et seulement si leur *pgcd* est 1.

Si on multiplie ou divise deux nombres par m, leur *pgcd* est multiplié ou divisé par m. Par conséquent, en divisant deux nombres par leur *pgcd*, on obtient deux nombres premiers entre eux.

▶ *La table des nombres premiers de 1 à 997.*

Nombres premiers de 1 à 997

1	2	3	5	7	11	13	17	19	23	29	31	37
41	43	47	53	59	61	67	71	73	79	83	89	97
101	103	107	109	113	127	131	137	139	149	151	157	163
167	173	179	181	191	193	197	199	211	223	227	229	233
239	241	251	257	263	269	271	277	281	283	293	307	311
313	317	331	337	347	349	353	359	367	373	379	383	389
397	401	409	419	421	431	433	439	443	449	457	461	463
467	479	487	491	499	503	509	521	523	541	547	557	563
569	571	577	587	593	599	601	607	613	617	619	631	641
643	647	653	659	661	673	677	683	691	701	709	719	727
733	739	743	751	757	761	769	773	787	797	809	811	821
823	827	829	839	853	857	859	863	877	881	883	887	907
911	919	929	937	941	947	953	967	971	977	983	991	997

Richard Colin

▼ *Le crible d'Ératosthène : méthode qui consiste à déterminer la table des nombres premiers.*

1	2	3	~~4~~	5	~~6~~	7	~~8~~	~~9~~	~~10~~	11	~~12~~	13	~~14~~	~~15~~
~~16~~	17	~~18~~	19	~~20~~	~~21~~	~~22~~	23	~~24~~	~~25~~	~~26~~	~~27~~	~~28~~	29	~~30~~
31	~~32~~	~~33~~	~~34~~	~~35~~	~~36~~	37	~~38~~	~~39~~	~~40~~	41	~~42~~	43	~~44~~	~~45~~
~~46~~	47	~~48~~	~~49~~	~~50~~	~~51~~	~~52~~	53	~~54~~	~~55~~	~~56~~	~~57~~	~~58~~	59	~~60~~
61	~~62~~	~~63~~	~~64~~	~~65~~	~~66~~	67	~~68~~	~~69~~	~~70~~	71	~~72~~	73	~~74~~	~~75~~
~~76~~	~~77~~	~~78~~	79	~~80~~	~~81~~	~~82~~	83	~~84~~	~~85~~	~~86~~	~~87~~	~~88~~	89	~~90~~
~~91~~	~~92~~	~~93~~	~~94~~	~~95~~	~~96~~	97	~~98~~	~~99~~	~~100~~	101	~~102~~	103	~~104~~	~~105~~
~~106~~	107	~~108~~	109	~~110~~	~~111~~	~~112~~	113	~~114~~	~~115~~	~~116~~	~~117~~	~~118~~	~~119~~	~~120~~
~~121~~	~~122~~	~~123~~	~~124~~	~~125~~	~~126~~	127	~~128~~	~~129~~	~~130~~	131	~~132~~	~~133~~	~~134~~	~~135~~
~~136~~	137	~~138~~	139	~~140~~	~~141~~	~~142~~	~~143~~	~~144~~	~~145~~	~~146~~	~~147~~	~~148~~	149	~~150~~
151	~~152~~	~~153~~	~~154~~	~~155~~	~~156~~	157	~~158~~	~~159~~	~~160~~	~~161~~	~~162~~	163	~~164~~	~~165~~
~~166~~	167	~~168~~	~~169~~	~~170~~	~~171~~	~~172~~	173	~~174~~	~~175~~	~~176~~	~~177~~	~~178~~	179	~~180~~
181	~~182~~	~~183~~	~~184~~	~~185~~	~~186~~	~~187~~	~~188~~	~~189~~	~~190~~	191	~~192~~	193	~~194~~	~~195~~
~~196~~	197	~~198~~	199	~~200~~	~~201~~	~~202~~	~~203~~	~~204~~	~~205~~	~~206~~	~~207~~	~~208~~	~~209~~	~~210~~
211	~~212~~	~~213~~	~~214~~	~~215~~	~~216~~	~~217~~	~~218~~	~~219~~	~~220~~	~~221~~	~~222~~	223	~~224~~	~~225~~

Palais de la Découverte, Paris

La propriété suivante est un des théorèmes fondamentaux de la théorie de la divisibilité : si un entier k divise le produit ab et si k est premier avec a, alors k divise b.

Pour le démontrer, supposons a, b, k non nuls. Par hypothèse, le *pgcd* de k et a est 1, il en résulte que le *pgcd* de kb et ab est b ; or k divise ab, k divise aussi kb, donc k divise le *pgcd* de kb et ab, et k divise b.

Ce résultat se généralise à un nombre quelconque de termes. En particulier, si k est premier avec a, il est aussi premier avec toute puissance de a.

Quant aux propriétés du *ppcm,* elles se déduisent de celles du *pgcd,* compte tenu du fait que :

$$ppcm\,(a, b) \cdot pgcd\,(a, b) = ab.$$

En effet, si $pgcd\,(a, b) = d$ alors $a = da'$, $b = db'$ avec a' et b' premiers entre eux. Soit m un multiple commun de a et b ; alors $m = ac = be = da'c = db'e$, d'où $a'c = b'e$; b' divisant $a'c$ et étant premier avec a' divise c ; on peut écrire

$$m = ac = da'c = da' \frac{c}{b'} b' = \frac{c}{b'} da'b'.$$

Tout multiple commun de a et b est multiple de $da'b'$; on en déduit que $ppcm\,(a, b) = da'b'$, d'où le résultat annoncé.

Les nombres premiers

Un nombre premier est un nombre qui n'a pas de diviseur propre (des diviseurs autres que lui-même et 1), par exemple 2, 3, 5, 7, 11, ...

Tout nombre entier peut être décomposé en produit de facteurs premiers, et cela de façon unique (à condition d'ordonner les facteurs),

$$a = p_1^{t_1} \cdot p_2^{t_2} \cdots p_r^{t_r}$$

où $p_1, ..., p_r$ sont des nombres premiers vérifiant

$$p_1 < p_2 < ... < p_r$$

et $t_1, ..., t_r$ des exposants entiers.

Cette décomposition est appelée *décomposition canonique*.

Supposons qu'il existe deux décompositions de a en facteurs premiers et groupons au début de chaque décomposition les éléments *communs* aux deux en un produit A.

Alors $a = AP = AQ$ implique $P = Q$. Soit alors p un des facteurs premiers intervenant dans le produit P, p est premier avec chacun des facteurs de Q (d'après la définition même de A), mais p divise Q (puisque p divise P), il y a donc contradiction, et l'hypothèse de l'existence de deux décompositions est fausse.

La décomposition canonique s'obtient en cherchant les facteurs premiers dans l'ordre croissant comme sur l'exemple suivant :

$n = 10\,800$

10 800	2
5 400	2
2 700	2
1 350	2
675	3
225	3
75	3
25	5
5	5
1	

$10\,800 = 2^4 \cdot 3^3 \cdot 5^2$.

Ceci va nous permettre de déterminer rapidement le *pgcd* ou le *ppcm* de deux nombres a et b, connaissant leur décomposition canonique. Soit $p_1, ..., p_r$ tous les facteurs premiers intervenant dans les deux décompositions (éventuellement sous la forme $p_i^\circ = 1$) avec les ordres de multiplicité (exposants) $t_1, ..., t_r$ pour a, $u_1, ..., u_r$ pour b ; si on désigne par m_i le plus petit des deux nombres t_i et u_i et par M_i le plus grand, alors :

$$pgcd\ (a, b) = p_1^{m_1} \cdot p_2^{m_2} \ldots p_r^{m_r}$$

$$ppcm\ (a, b) = p_1^{M_1} \cdot p_2^{M_2} \ldots p_r^{M_r}$$

La répartition des nombres premiers est un problème posé dès l'Antiquité grecque et qui reste encore ouvert. Euclide, déjà, a montré l'existence d'une infinité de nombres premiers grâce à l'argument suivant : soit $p_1, p_2, ..., p_k$, k nombres premiers ; alors $n = (p_1 \cdot p_2 \ldots p_k) + 1$ est premier avec $p_1 \cdot p_2 \ldots p_k$, car si un nombre divisait à la fois le produit $p_1 \cdot p_2 \ldots p_k$ et n, il devrait diviser 1.

D'autre part, n admet au moins un facteur premier, ce facteur est nécessairement distinct des $p_1, ..., p_k$; il existe donc un nombre premier distinct des k nombres premiers considérés.

La distribution des nombres premiers dans l'ordre croissant des entiers est extrêmement irrégulière, et a fait l'objet de recherches nombreuses et très approfondies. Si l'on désigne par $n!!$ le produit des nombres premiers inférieurs ou égaux à n : $2!! = 2$, $3!! = 4!! = 6$, $5!! = 6!! = 30$, $7!! = 8!! = 9!! = 10!! = 210$ $11!! = 12!! = 2\,310$, $13!! = 14!! = 15!! = 16!! = 30\,030$, etc.
alors $n!! + 1$ peut être premier ou non mais

$$n!! + 2, ..., n!! + n$$

sont tous des nombres composés (non premiers).

Par suite, on peut trouver dans la liste des nombres premiers des trous de longueur arbitraire, par exemple deux nombres premiers consécutifs dont la différence est au moins égale à 100, il suffit de remarquer que $101!! + 2$, ..., $101!! + 101$ est suite de 100 nombres composés. Mais on peut aussi rencontrer plusieurs nombres premiers très proches, par exemple 101, 103, 107, 109. Cependant, les nombres premiers se raréfient à mesure qu'ils deviennent de plus en plus grands.

On démontre que, si l'on désigne par $\pi(n)$ le nombre de nombres premiers inférieurs ou égaux à n, on a la relation :

$$\lim_{n \to \infty} \frac{\pi(n)}{n} = 0 \left(\text{plus précisément } \pi(n) \text{ tend vers } \frac{n}{\operatorname{Log} n}\right)$$

D'autre part, les nombres premiers se raréfient plus lentement que les carrés des nombres entiers ; la série $\sum_{n=1}^{\infty} \frac{1}{n^2}$ est convergente alors que la série $\sum_{n=1}^{\infty} \frac{1}{p_n}$ des inverses des nombres premiers ordonnés en une suite croissante $p_1, ..., p_n, ...$ est divergente.



Palais de la Découverte, Paris

▲ ***Le mathématicien* Pierre de Fermat *(1601-1665) célèbre pour ses nombreuses découvertes en arithmétique. Ses principaux écrits ont été publiés par son fils, en 1679, sous le titre de* Varia Opera mathematica.**

L'hypothèse de Fermat

Pierre de Fermat (1601-1665), conseiller au Parlement de Toulouse, est célèbre pour ses nombreuses découvertes en arithmétique. Reprenant un problème posé par Diophante d'Alexandrie au IVe siècle : trouver tous les triangles rectangles dont les trois côtés sont mesurés par des nombres entiers, c'est-à-dire, résoudre dans ℕ l'équation :

$$x^2 + y^2 = z^2$$

dont la solution est donnée par

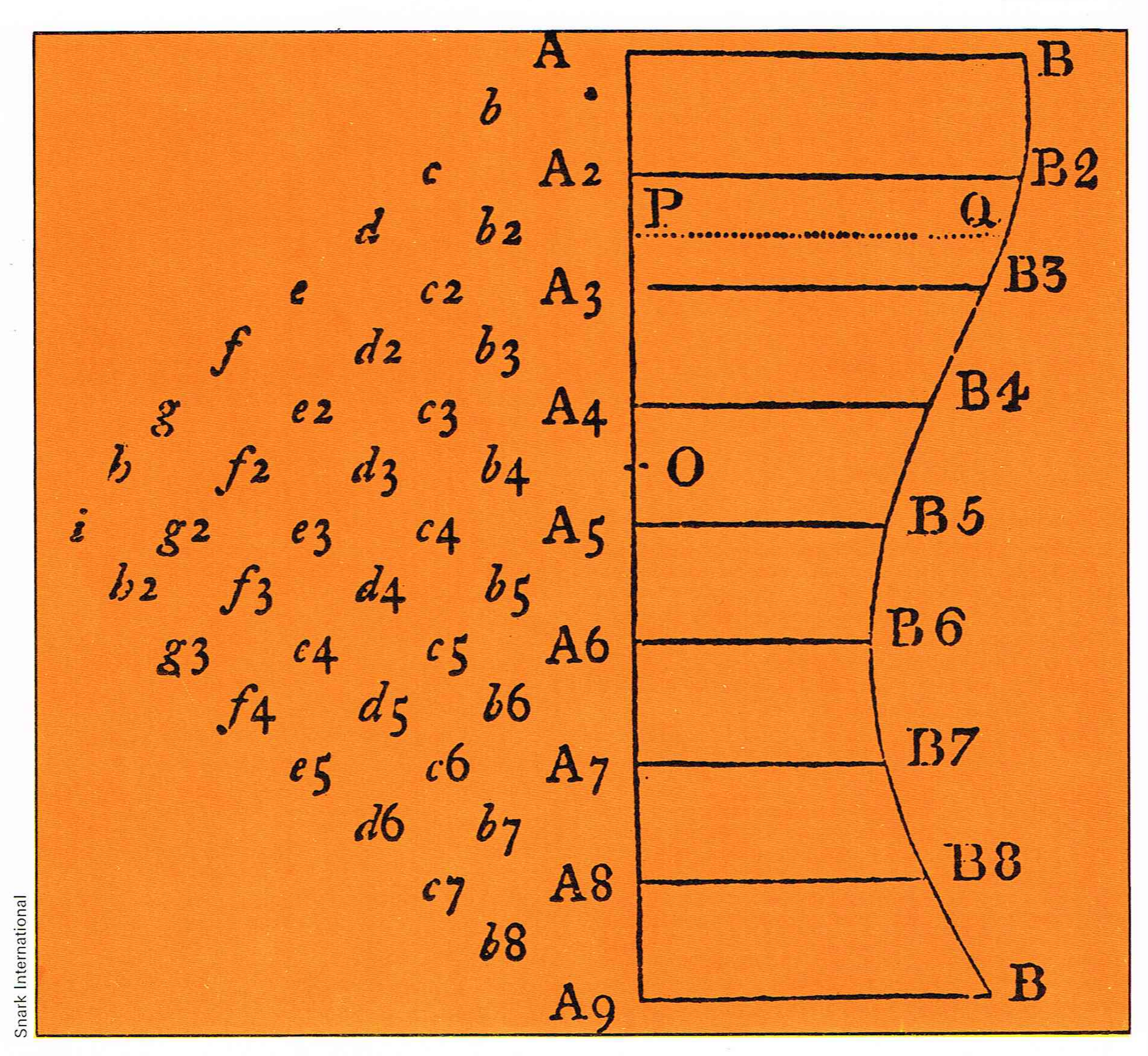
$$y = ab, \quad z = \frac{a^2 + b^2}{2}, \quad x = \frac{a^2 - b^2}{2}$$

Fermat déclare avoir découvert une méthode remarquable prouvant que l'équation :

$$x^n + y^n = z^n$$

n'a aucune solution en nombres entiers pour n supérieur à 2. Malheureusement, il ne laissa pas de trace de sa démonstration. Des travaux ultérieurs, dus en particulier à Kummer au XIXe siècle, montrèrent que l'hypothèse de Fermat est exacte pour de nombreuses valeurs de n, mais la démonstration complète n'en a pas encore été élaborée.

▶ *Figure arithmétique extraite des Œuvres de Newton (édition du XIXe siècle).*



Snark International

▶ *Page ci-contre, document ancien extrait d'un ouvrage maghrébin et utilisant dans un « dessin magique » des symboles numériques (musée du Caire).*

Les nombres rationnels

De même que l'équation $a + x = b$ n'a pas toujours de solution dans $\mathbb{N}$, l'équation $ax = b$ n'a pas de solution dans $\mathbb{N}$, ni dans $\mathbb{Z}$ pour toutes les valeurs du couple (a, b). Cela va nous amener à définir un nouvel ensemble de nombres, $\mathbb{Q}$, ensemble des nombres rationnels, dans lequel l'équation $ax = b$ peut être résolue pour tout a $(\neq 0)$ et tout b.

Nous construirons cet ensemble $\mathbb{Q}$ par un procédé analogue à celui que nous avons utilisé pour $\mathbb{Z}$. La relation que nous allons considérer dans $\mathbb{Z} \times \mathbb{Z}^*$ $(\mathbb{Z}^* = \mathbb{Z} - \{0\})$ est la suivante :

$$(a, b) \sim (c, d) \text{ si, et seulement si } ad = bc.$$

On vérifie sans peine que cette relation est une relation d'équivalence, ce qui permet d'obtenir une partition de $\mathbb{Z} \times \mathbb{Z}^*$ en classes d'équivalence deux à deux disjointes.

Nous désignerons par $\frac{a}{b}$ la classe à laquelle appartient le couple (a, b) et par $\mathbb{Q}$ l'ensemble des classes d'équivalence :

$$\mathbb{Q} = \left\{ \frac{a}{b} \,\middle|\, a \in \mathbb{Z}, b \in \mathbb{Z}^* \right\}$$

Il reste encore à définir sur $\mathbb{Q}$ une addition et une multiplication qui prolongent celles que l'on connaît dans $\mathbb{Z}$; pour cela nous poserons :

$$\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd}$$

$$\frac{a}{b} \cdot \frac{c}{d} = \frac{ac}{bd}$$

ce qui, *a priori*, a un sens puisque $bd \neq 0$ (la classe $\frac{a}{b}$ n'est définie que pour $b \neq 0$).

Considérons dans $\mathbb{Q}$ le sous-ensemble des classes de la forme $\left\{ \frac{a}{1} \,\middle|\, a \in \mathbb{Z} \right\}$; ce sous-ensemble est isomorphe à $\mathbb{Z}$; en effet l'application φ qui, à l'élément a de $\mathbb{Z}$ associe la classe $\frac{a}{1}$, est bijective (elle est manifestement surjective, et si $\frac{a}{1} = \frac{b}{1}$, alors $(a, 1) \sim (b, 1)$, d'où $a = b$, ce qui prouve qu'elle est injective) et vérifie :

$$\varphi\,(a + b) = \frac{a + b}{1} = \frac{a}{1} + \frac{b}{1} = \varphi\,(a) + \varphi\,(b)$$

$$\varphi\,(a \cdot b) = \frac{a \cdot b}{1} = \frac{a}{1} \cdot \frac{b}{1} = \varphi\,(a) \cdot \varphi\,(b).$$

Par suite, on identifiera $\frac{a}{1}$ et a (comme on l'avait fait pour $[0, n]$ et n), ce qui permettra de considérer $\mathbb{Z}$ comme un sous-ensemble de $\mathbb{Q}$.

$\mathbb{Q}$ possède toutes les propriétés de $\mathbb{Z}$; comme $\mathbb{Z}$, $\mathbb{Q}$ est un anneau commutatif, unitaire, intègre ($c \neq 0$, $ac = bc \Rightarrow a = b$), mais en outre dans $\mathbb{Q}$ l'équation $ax = b$ admet une solution unique pour tout couple (a, b) $(a \neq 0)$.

En effet, posons $a = \frac{a_1}{a_2}$, $x = \frac{x_1}{x_2}$, $b = \frac{b_1}{b_2}$

($a_1 \neq 0$ car $a \neq 0$; $b_2 \neq 0$), alors $ax = b$ équivaut à $\frac{a_1\,x_1}{a_2\,x_2} = \frac{b_1}{b_2}$ ou encore $(a_1\,b_2)\,x_1 = (a_2\,b_1)\,x_2$ par définition de l'égalité entre deux rationnels ; il en résulte $\frac{x_1}{x_2} = \frac{a_2\,b_1}{a_1\,b_2}$;

la solution existe ($a_1 b_2 \neq 0$) ; elle est unique, on la désigne par $\frac{b}{a}$.

L'ensemble $\mathbb{Q}^* = \mathbb{Q} - \{0\}$ des nombres rationnels non nuls est un groupe commutatif pour la multiplication. Il suffit de vérifier que tout rationnel α non nul admet un symétrique tel que $\alpha\alpha' = \alpha'\alpha = 1$, ce qui est une conséquence immédiate de la résolution dans l'équation $ax = b$. On aura $\alpha' = \frac{1}{\alpha}$.

On en déduit de façon immédiate la régularité de la multiplication dans $\mathbb{Q}$. Supposons $\alpha\, x = \alpha y$, $\alpha \neq 0$; alors

$$\frac{1}{\alpha}\,(\alpha\, x) = \frac{1}{\alpha}\,(\alpha\, y) \text{ et } x = y.$$

Un anneau commutatif qui est aussi un groupe pour la multiplication est un *corps (cf. Structures)*. Si la deuxième loi est commutative, le *corps est commutatif*.

Ainsi $\mathbb{Q}$ est un corps commutatif. C'est le plus petit des corps commutatifs contenant $\mathbb{Z}$ dans lequel l'équation $ax = b$ a une solution pour tout a ($\neq 0$) et tout b appartenant à $\mathbb{Z}$.

Nous allons maintenant montrer comment s'étend à $\mathbb{Q}$ l'ordre naturel de $\mathbb{Z}$.

Pour tout couple d'entiers a, b avec $b \neq 0$, $\frac{a}{b} \geqslant 0$ si, et seulement si $ab \geqslant 0$ et pour deux nombres rationnels quelconques α, β, $\alpha \leqslant \beta$ si, et seulement si $\alpha - \beta \geqslant 0$.

Alors les propriétés **(01′)**, **(02′)**, **(03′)** du paragraphe précédent sont vérifiées.

On montre aisément que ces conditions sont équivalentes à $\frac{a}{b} \geqslant 0$ si, et seulement si $ab \geqslant 0$ ($a \in \mathbb{Q}, b \in \mathbb{Q}^*$).

L'ordre ainsi défini dans $\mathbb{Q}$ prolonge l'ordre obtenu dans $\mathbb{Z}$.

Le corps $\mathbb{Q}$ satisfait à la propriété suivante : « Pour tout couple (a, b) de nombres rationnels strictement positifs, il existe un entier n tel que $b \geqslant na$. » Propriété que l'on traduit en disant que $\mathbb{Q}$ est un *corps archimédien*.

Combien y a-t-il d'éléments dans $\mathbb{Q}$? Nous allons voir qu'il y en a exactement autant que dans $\mathbb{N}$, résultat qui peut paraître paradoxal si l'on oublie que $\mathbb{N}$, comme $\mathbb{Q}$, n'est pas un ensemble fini. L'application de $\mathbb{Q}$ dans $\mathbb{N} \times \mathbb{N}^*$ qui à $\frac{a}{b}$ associe le couple (a, b) est manifestement bijective, donc Card $\mathbb{Q}$ = Card ($\mathbb{N} \times \mathbb{N}^*$).

Les éléments de $\mathbb{N} \times \mathbb{N}^*$ peuvent être rangés de la façon suivante :

(0, 1) → (0, 2)   (0, 3) → (0, 4) ...
(1, 1)   (1, 2)   (1, 3)   (1, 4) ...
(2, 1)   (2, 2)   (2, 3)   (2, 4) ...
(3, 1)   (3, 2)   (3, 3)   (3, 4) ...
(4, 1)   (4, 2)   (4, 3) ...

Il est possible de les numéroter en suivant les flèches (procédé diagonal) ; on déduit que

$$\text{Card } \mathbb{N} \times \mathbb{N}^* = \text{Card } \mathbb{N},$$

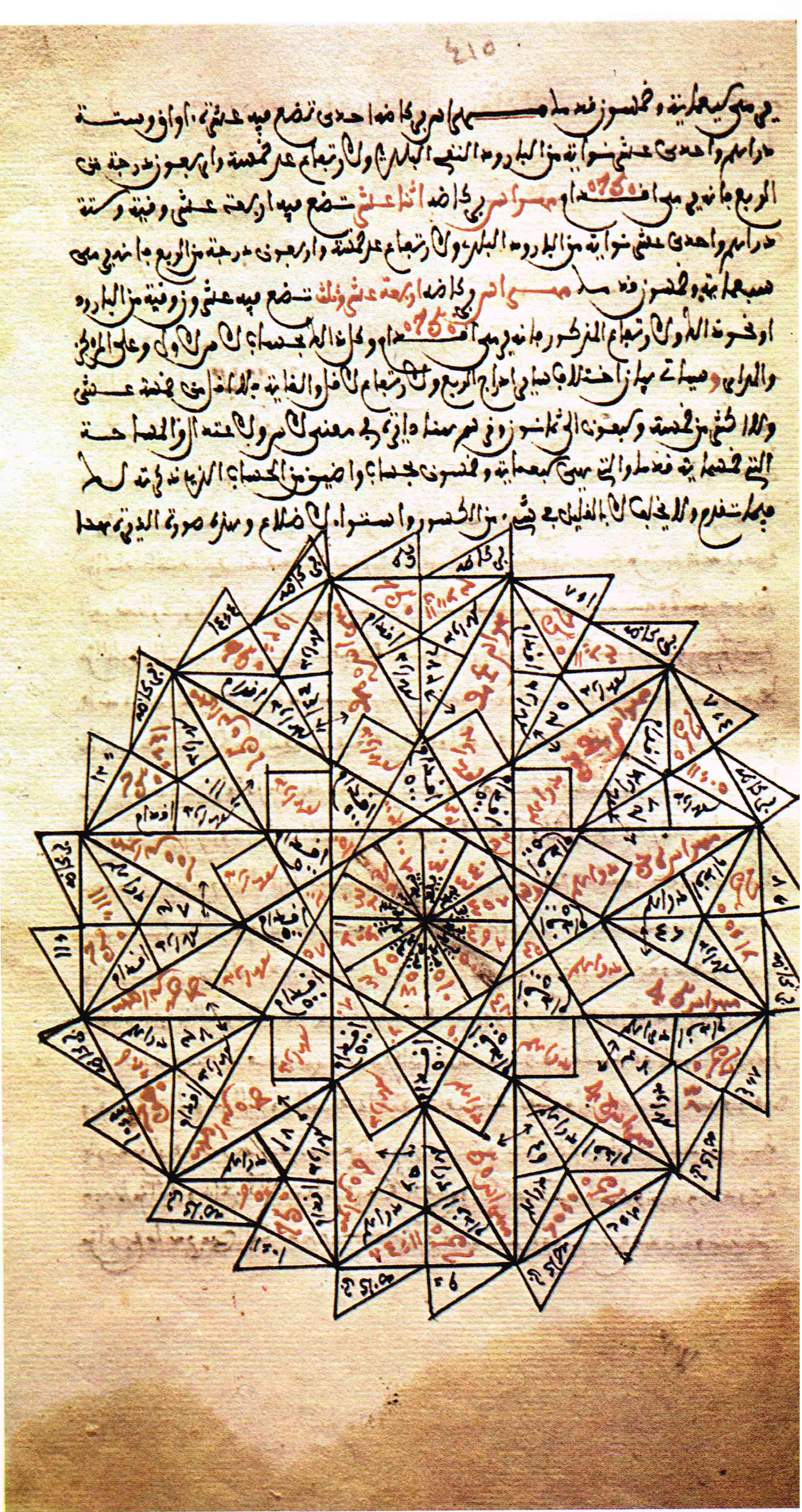
d'où le résultat annoncé : Card $\mathbb{Q}$ = Card $\mathbb{N}$, ce qui s'énonce encore :

$\mathbb{Q}$ est dénombrable.

Les nombres réels

Soit $\sqrt{2}$ le nombre dont le carré est égal à 2. Nous allons montrer que $\sqrt{2}$ ne peut pas être rationnel. En effet, supposons qu'il existe un nombre rationnel $\frac{p}{q}$, avec p et q premiers entre eux, tel que $\frac{p}{q} = \sqrt{2}$, alors $2 = \frac{p^2}{q^2}$ et $p^2 = 2\,q^2$.

2 divise $2\,q^2$, donc 2 divise p^2 et 2 divise p (il suffit de considérer la décomposition de p en facteurs premiers) ; si 2 divise p, 4 divise p^2 donc $2\,q^2$; donc 2 divise q^2 ce qui contredit que p et q soient premiers entre eux puisque 2 divise également p^2 ; l'hypothèse $\sqrt{2}$ rationnel



R. Michaud - Rapho

est donc illégitime. Nous venons ainsi de mettre en évidence la nécessité de construire un ensemble plus vaste que $\mathbb{Q}$, permettant notamment de résoudre des équations du type $x^2 = 2$. Cet ensemble sera l'ensemble des nombres réels $\mathbb{R}$.

L'ensemble $\mathbb{R}$ a pour caractéristique essentielle d'être un corps continu.

Un corps ordonné K est dit continu s'il satisfait à la condition suivante, appelée postulat de continuité de Dedekind :

Étant donné une partition des éléments de K en deux classes non vides et séparées, c'est-à-dire telles que tout élément de la première classe soit inférieur (au sens de l'ordre de K) à tout élément de la seconde, il existe dans K un et un seul élément séparateur (coupure) des deux classes (ce qui revient à dire un, et un seul, élément s tel que tout élément de la première classe ne soit pas supérieur à s, et que tout élément de la seconde ne soit pas inférieur à s).

Une des méthodes de construction des réels consiste à considérer l'ensemble formé de :

1) tous les nombres rationnels ;

2) toutes les partitions de l'ensemble $\mathbb{Q}$ des nombres rationnels en classes non vides et séparées dont la coupure n'appartient pas à $\mathbb{Q}$.

On identifie un nombre rationnel quelconque α à la partition de $\mathbb{Q}$ en deux classes non vides et séparées par α, précisant que α appartient à la seconde classe. Alors il est aisé de définir sur $\mathbb{R}$ une addition, une multiplication et une relation d'ordre.

Nous nous limiterons à montrer comment on définit l'addition.

Soit A et B deux partitions de $\mathbb{Q}$ en deux classes non vides et séparées (A_1, A_2) et (B_1, B_2). Par A + B nous désignerons la partition de $\mathbb{Q}$ qui admet comme première classe l'ensemble des nombres rationnels de la forme $a_1 + b_1$, où a_1 appartient à A_1 et b_1 appartient à B_1, et comme seconde classe tous les autres nombres rationnels.

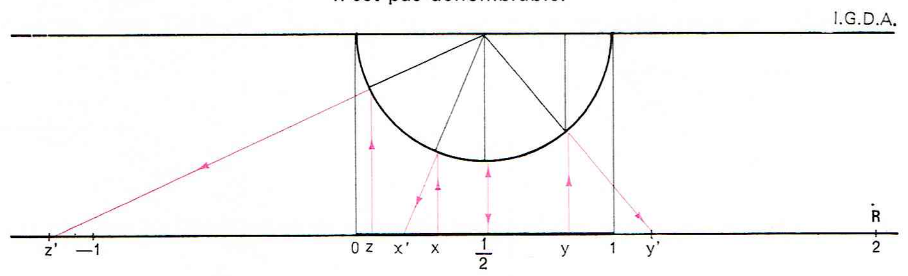
Le nombre réel A + B ainsi défini est appelé somme des nombres réels A et B. On démontrerait sans peine que toutes les propriétés de l'addition dans $\mathbb{Q}$ sont encore vraies dans $\mathbb{R}$.

Le même procédé nous permettra de définir le produit AB de deux nombres réels et d'étendre à $\mathbb{R}$ l'ordre défini sur $\mathbb{R}$, faisant de $\mathbb{R}$ un corps continu, extension ordonnée du corps des rationnels $\mathbb{Q}$.

Si nous reprenons l'exemple cité plus haut, on définira $\sqrt{2}$ à l'aide de la partition de $\mathbb{Q}$ en deux classes non vides et séparées : celle des rationnels dont le carré est inférieur ou égal à 2, celle des rationnels dont le carré est supérieur ou égal à 2. Il existe dans $\mathbb{R}$ un et un seul élément qui sépare ces deux classes ; on le note $\sqrt{2}$.

L'ensemble $\mathbb{R}$ des nombres réels n'est pas dénombrable. $\mathbb{R}$ peut être mis en correspondance biunivoque avec l'intervalle $]0, 1[= \{x \mid x \in \mathbb{R}, 0 < x < 1\}$, comme l'illustre la figure ci-dessous ; il suffit donc de démontrer que $]0, 1[$ n'est pas dénombrable.

I.G.D.A.



▲ ***Correspondance biunivoque de l'ensemble] 0,1 [avec l'ensemble $\mathbb{R}$ des nombres réels.***

Tout nombre réel de l'intervalle $]0, 1[$ admet une représentation décimale unique $x = 0, x_1 x_2 \ldots, x_n \ldots$, où $x_1, x_2, \ldots, x_n, \ldots$ sont des entiers non tous nuls et non tous égaux à 9 à partir d'un certain rang. Réciproquement, tout développement décimal de ce type représente un élément bien déterminé de $]0, 1[$. Il sera suffisant, pour notre propos, de montrer que l'ensemble de ces développements décimaux n'est pas dénombrable. Nous le démontrerons par l'absurde à l'aide du procédé diagonal de Cantor.

Supposons que l'ensemble des développements décimaux des nombres réels de l'intervalle $]0, 1[$ soit dénombrable ; alors on peut tous les numéroter par l'intermédiaire d'une suite $a_1, a_2, \ldots, a_n$, de nombres compris entre 0 et 1 ; $a_1, \ldots, a_n, \ldots$, admettent des développements décimaux :

$$\begin{aligned} a_1 &= 0, a_{11}\, a_{12} \ldots a_{1n} \ldots \\ a_2 &= 0, a_{21}\, a_{22} \ldots a_{2n} \ldots \\ &\ldots\ldots\ldots\ldots \\ a_n &= 0, a_{n1}\, a_{n2} \ldots a_{nn} \ldots \\ &\ldots\ldots\ldots\ldots \end{aligned}$$

Le développement décimal infini $b = 0, b_1 b_2 \ldots b_n \ldots$, défini pour tout n, par $b_n \neq 0$, $b_n \neq 9$, $b_n \neq a_{nn}$ (par exemple $b_n = 2$ si $a_{nn} = 1$, $b_n = 1$ si $a_{nn} \neq 1$) est du type considéré ; il devrait appartenir à la suite $a_1, a_2, \ldots, a_n, \ldots$, ce qui est absurde, car b diffère de a_1 pour le premier chiffre après la virgule, de a_2 pour le second chiffre, de a_n pour le n-ième chiffre.

L'hypothèse faite est donc fausse, et on peut affirmer que $\mathbb{R}$ n'est pas dénombrable.

Comme $\mathbb{Q}$ est dénombrable, il existe une infinité non dénombrable d'irrationnels ; en effet si $\mathbb{R} - \mathbb{Q}$ était dénombrable, $\mathbb{R}$ le serait.

Les nombres algébriques

Parmi les nombres irrationnels, on considère un sous-ensemble, celui des *nombres algébriques réels* : on appelle ainsi tout nombre réel a qui est racine d'un polynôme

$$P(X) = a_n X^n + a_{n-1} X^{n-1} + \ldots + a_1 X + a_0$$

dont les coefficients sont des entiers de signes quelconques.

Ainsi $\sqrt{2}$ qui est racine du polynôme $X^2 - 2$ est un nombre algébrique. Tout nombre rationnel $\frac{p}{q}$ $(q \neq 0)$ est un nombre algébrique, $\frac{p}{q}$ est racine de $qX - p = 0$.

Si $a_n = 1$, a est un *entier algébrique réel.* Tout nombre algébrique est le quotient d'un entier algébrique réel par un entier. Pour le constater, il suffit de poser $X = \frac{x}{a_n}$ et de multiplier le polynôme P par $(a_n)^{n-1}$, alors

$$P(x) = x^n + a_{n-1}\, x^{n-1} + a_n \cdot a_{n-2} \cdot x^{n-2} + a_n \cdot a_0$$

L'ensemble des nombres réels algébriques est dénombrable, comme le montre le raisonnement suivant :

On remarque que tout nombre algébrique a est racine d'un polynôme P (X) et d'un seul vérifiant : les coefficients de P (X) sont des entiers rationnels premiers entre eux dans leur ensemble (sinon on divise tous les coefficients par leur *pgcd*), P (X) est irréductible dans $\mathbb{Z}$ (si P admet une racine entière α, d'ordre n, on considère le polynôme P' obtenu en simplifiant P par $(X - \alpha)^n$). Appelons degré de a le degré de P (X), et hauteur de a le maximum des valeurs absolues des coefficients de P (X).

Pour d, h entiers positifs fixés, $E_{h,\,d}$ désigne l'ensemble des entiers algébriques de degré d et de hauteur h. $E_{h,\,d}$ est manifestement un ensemble fini. D'autre part, comme l'ensemble $\mathbb{N} \times \mathbb{N}$ est dénombrable, alors la réunion des ensembles $E_{h,\,d}$ pour toutes les valeurs possibles de h et de d est dénombrable, donc aussi l'ensemble des nombres réels algébriques.

Il existe donc dans $\mathbb{R}$ des nombres qui ne sont pas algébriques : ce sont les *nombres transcendants.*

Les nombres transcendants

Liouville donna en 1851 un procédé de construction de nombres transcendants :

soit a un nombre réel tel que, pour tout entier n, il existe une infinité de fractions irréductibles $\frac{p}{q}$ pour lesquelles $\left| a - \frac{p}{q} \right| < \frac{1}{q^n}$.

On démontre assez facilement qu'un tel nombre est transcendant, et on peut construire des nombres a satisfaisant à une telle condition en prenant des sommes de séries très rapidement convergentes à termes rationnels, par exemple :

$$a = \sum_{n=1}^{\infty} \left(\frac{1}{2}\right)^{n^n}$$

Le problème de savoir si un nombre donné est ou non transcendant est souvent fort difficile à résoudre et fait appel à des résultats assez élaborés d'analyse.

Hermite (1822-1901) a démontré que $e = 2{,}718...$, base des logarithmes népériens, est un nombre transcendant. Plus tard, Lindemann (1852-1939) a établi la transcendance de $\pi = 3{,}141\,59...$

Les méthodes actuellement employées sont liées à l'approximation, par des nombres rationnels, des nombres dont on cherche à démontrer la transcendance. Mais jusqu'à présent, on n'a pas pu établir la transcendance ou l'irrationalité de la constante d'Euler :

$$C = \lim_{n \to \infty} \left(1 + \frac{1}{2} + ... + \frac{1}{n} - \text{Log } n\right)$$

définie par des séries qui convergent très lentement.

L'approximation des irrationnels par les rationnels est d'un grand intérêt ; c'est l'objet principal de la *théorie des approximations diophantiennes* (du nom de Diophante d'Alexandrie, mathématicien du IVe siècle). Dans le cas d'un seul irrationnel, les *fractions continuées* jouent un rôle essentiel (leurs propriétés servirent à Huyghens, dès 1650, pour le calcul des engrenages des horloges astronomiques), mais les résultats déterminants pour cette théorie furent établis par Liouville en 1884 et par Roth en 1955.

Les nombres complexes

Dès le XVIe siècle, les algébristes italiens Tartaglia (1506-1559), Cardan (1501-1576) et Ferrari (1522-1565) eurent l'idée audacieuse de désigner par le symbole $\sqrt{-1}$ la racine carrée apparemment inexistante de -1.

Ils purent alors résoudre non seulement les équations du 2e degré, mais aussi celles du 3e et même du 4e. Malgré les travaux importants de D'Alembert, puis de Gauss, il faudra attendre le XVIIIe siècle et les publications d'Euler pour voir disparaître chez les mathématiciens toute méfiance à l'égard des nombres « imaginaires ».

L'impossibilité de résoudre dans $\mathbb{R}$ l'équation $x^2 + 1 = 0$ suggère l'idée de créer un ensemble de nombres plus vaste. Cependant, dans un premier temps, cela peut paraître impossible, car jamais dans un corps ordonné on ne pourra extraire la racine carrée d'un nombre négatif. Extraire la racine carrée d'un tel nombre $-a$ équivaut en fait à trouver une racine du polynôme $x^2 + a$ qui est strictement positif comme somme d'un nombre positif a et d'une quantité non négative x^2.

La racine carrée d'un nombre négatif ne pourra appartenir à aucun corps ordonné qui soit une extension de $\mathbb{R}$.

On démontre même plus : aucun corps ordonné, construit comme une extension ordonnée de $\mathbb{R}$, ne possède de racine d'un polynôme réel (à coefficients dans $\mathbb{R}$), irréductible dans $\mathbb{R}$.

L'adjonction à $\mathbb{R}$ d'une racine carrée d'un nombre réel négatif, ou plus généralement d'une racine d'un polynôme réel irréductible dans $\mathbb{R}$, pourra être faite à condition de renoncer à l'ordre. Le nouvel ensemble ainsi obtenu ne pourra être totalement ordonné.

Une des extensions les plus importantes de ce type va être le *corps des complexes*, obtenu par adjonction à $\mathbb{R}$ d'une racine du polynôme $x^2 + 1$.

Divers procédés permettent de construire le corps des complexes; nous en décrirons un des plus simples. Comme le calcul complexe opère sur des expressions du type $a + bi$ où a et b sont des nombres réels et i une racine du polynôme $x^2 + 1$, nous définirons sur l'ensemble $\mathbb{C}'$ des couples de nombres réels (a, b) l'addition et la multiplication suivantes :

(1) $(a, b) + (c, d) = (a + b, c + d)$

(2) $(a, b)\,(c, d) = (ac - bd, ad + bc)$

$\mathbb{C}'$ muni de ces deux opérations est un corps.

Nous nous limiterons à déterminer le symétrique par rapport à la multiplication de tout couple (a, b) différent de $(0, 0)$, élément neutre pour l'addition ; ce sera :

$$\left(\frac{a}{a^2 + b^2}, \frac{-b}{a^2 + b^2}\right)$$

car

$$(a, b) \cdot \left(\frac{a}{a^2 + b^2}, \frac{-b}{a^2 + b^2}\right) =$$

$$\left(\frac{a^2}{a^2 + b^2} + \frac{b^2}{a^2 + b^2}, \frac{-ab}{a^2 + b^2} + \frac{ab}{a^2 + b^2}\right) = (1, 0)$$

où $(1, 0)$ est l'élément neutre de la multiplication.

La correspondance qui associe à tout élément a de $\mathbb{R}$ le couple $(a, 0)$ est un isomorphisme de $\mathbb{R}$ dans le sous-ensemble de $\mathbb{C}'$ formé par les couples $(a, 0)$. Cette correspondance est en effet manifestement bijective et telle que

$$(a, 0) + (b, 0) = (a + b, 0)$$

$$(a, 0)\,(b, 0) = (ab, 0)$$

compte tenu des identités **(1)** et **(2)** précédentes.

Si on identifie alors le couple $(a, 0)$ de $\mathbb{C}'$ et l'élément a de $\mathbb{R}$, ce qui peut se faire tout en respectant l'addition et la multiplication dans $\mathbb{C}'$, on obtient le corps $\mathbb{C}$, extension du corps $\mathbb{R}$.

Tout élément de $\mathbb{C}$ peut se mettre sous la forme $a + bi$ où $i = (0, 1)$. C'est une conséquence immédiate de l'identité :

$$(a, b) = (a, 0) \cdot (1, 0) + (b, 0)\,(0, 1)$$

Dans $\mathbb{C}$, on aura les égalités suivantes :

$$a + bi + c + di = a + c + (b + d)\,i$$

$$(a + bi) \cdot (c + di) = ac - bd + (ad + bc)\,i$$

$$(a + bi)\,(a - bi) = a^2 + b^2 \qquad (i\,(-i) = 1)$$

De la dernière égalité, il résulte $i^2 = -1$, ce qui n'est pas surprenant, car i (comme aussi $-i$) a été choisi comme racine du polynôme $x^2 + 1 = 0$.

Les règles de calcul sur les nombres complexes $z = a + ib$ sont les mêmes que celles que l'on connaît dans $\mathbb{R}$, compte tenu du fait que l'on pourra remplacer i^2 par -1 dès que cette quantité apparaîtra, ce qui rend le calcul dans $\mathbb{C}$ très opérationnel.

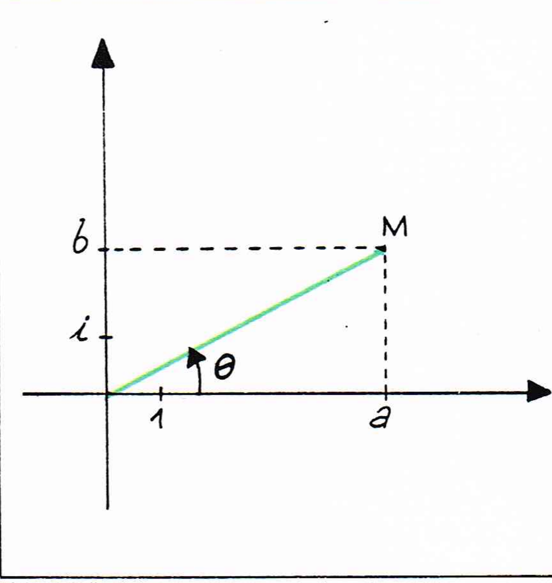
Si un corps $\mathbb{C}''$ contient une racine j du polynôme $x^2 + 1$, alors l'ensemble des nombres $a + bj$ constitue un sous-corps de $\mathbb{C}''$ isomorphe à $\mathbb{C}$. C'est le plus petit sur-corps de $\mathbb{R}$ dans lequel $x^2 + 1$ est réductible.

En construisant $\mathbb{C}$ comme une extension algébrique de $\mathbb{R}$ dans laquelle le polynôme $x^2 + 1$ admet une racine, on obtient un résultat beaucoup plus fort, connu sous le nom de *théorème de D'Alembert :* « Dans $\mathbb{C}$ tout polynôme non constant admet au moins une racine. »

Il en résulte que tout polynôme de degré n peut se décomposer dans $\mathbb{C}$ en produit de n facteurs du premier degré. On traduit cette propriété en disant que $\mathbb{C}$ est algébriquement clos.

En particulier, tout nombre complexe a n racines n-ièmes ; en effet, à tout nombre complexe α on peut associer n nombres complexes x tels que $x^n = \alpha$.

La représentation géométrique des nombres complexes va être la source de nombreuses applications, notamment pour les transformations géométriques et la trigonométrie (schéma ci-dessous).



Richard Colin

◀ ***On appelle affixe du nombre complexe*** **z = a + ib** ***le point M de coordonnées (a, b) dans un repère orthonormé ; alors*** $z = a + ib = \rho\,(\cos\theta + i\sin\theta) = \rho \cdot e^{i\theta}$ ***(pour cette dernière égalité, se reporter au chapitre*** **Trigonométrie*).***

▶ *Page ci-contre, un ruban perforé pour ordinateur. Les caractères numériques sont représentés dans le* système binaire.

▼ *Figuration du nombre, nombres figurés et calcul concret : exemples de figurations digitales numériques extraits de* Sùmma, *œuvre de* Pacioli, *Venise, 1494.*

Les quaternions

On généralise les nombres complexes en nombres hypercomplexes de la forme :

$$a = b_0 e_0 + ... + b_{n-1} e_{n-1}$$

où les b sont des nombres réels, les e des nombres complexes, appelés unités du système considéré et vérifiant :

$$e_j e_k = \sum_{s=0}^{n-1} \gamma_{jks} e_s.$$

On obtient divers systèmes suivant les valeurs que l'on donne aux γ_{jks}.

Les nombres complexes ordinaires sont d'ordre 2 et vérifient :

$$e_0 e_0 = e_0, \quad e_0 e_1 = e_1 e_0 = e_1, \quad e_1 e_1 = - e_0$$

$$(e_0 = 1, \quad e_1 = i)$$



Palais de la Découverte, Paris

Le plus célèbre des systèmes d'ordre 4 est celui des quaternions, dû à Hamilton et vérifiant :

$$e_j e_j = - e_0 \qquad e_j e_k = + e_h$$

$$e_k e_j = - e_k$$

où j, h, k est une permutation circulaire des nombres 1, 2, 3.

Systèmes de numération

Les entiers

Toutes les constructions que nous venons de décrire ne dépendaient pas de la manière de représenter les nombres naturels, entiers, rationnels, réels. Cependant, cette représentation a une grande importance dans tous les problèmes d'arithmétique appliquée, que ce soit dans la vie quotidienne ou dans la recherche scientifique et technique.

Le système actuel de *représentation décimale* des nombres ou *système décimal,* ou encore *système à base dix,* d'origine indienne, est parvenu en Europe à la fin du XIIe siècle, grâce au rayonnement mathématique arabe, par l'intermédiaire de Leonardo Filonacci, auteur du *Liber abaci* (1202). Il en a résulté un progrès décisif par rapport aux systèmes de numération grec et romain.

Dans le système de numération décimale, comme dans tout système de numération positionnelle, les nombres entiers positifs sont représentés à l'aide d'une suite ordonnée finie d'entiers non négatifs tous strictement inférieurs à un nombre fixé à l'avance qui sera *la base* du système, 10 dans le cas de la représentation décimale.

L'usage du signe —, de la virgule et d'une suite infinie d'entiers inférieurs à la base permet alors de représenter tout nombre réel. Pour les nombres réels ainsi représentés, les règles d'addition et de multiplication seront très simples et faciliteront le calcul, même pour des nombres très grands. Elles seront valides quelle que soit la base, en particulier en base 2 qui est d'une très grande utilité pour le calcul automatique *(cf. Informatique).*

Tout entier positif peut s'écrire (grâce aux divisions avec reste suivant les puissances décroissantes de b) d'une façon et d'une seule sous la forme :

$$a = q_m b^m + q_{m-1} b^{m-1} + ... + q_1 b + q_0$$

où q_i est un entier positif vérifiant pour tout i $0 \leqslant q_i < b$.

A une représentation de ce type, on peut faire correspondre biunivoquement la suite finie $q_m q_{m-1} ... q_1 q_0$. D'où, la base b étant fixée, une correspondance biunivoque entre un nombre entier positif quelconque et la suite $q_m q_{m-1} ... q_1 q_0$ (suite qui ne commence jamais par zéro). Dans le cas $b = 10$, les q_i sont 0, 1, ..., 9 ; on obtient la représentation décimale usuelle.

Exemple : $9\,326 = 9 10^3 + 3 10^2 + 2 10 + 6.$

Si on veut représenter un nombre en base 12, il est nécessaire d'introduire deux symboles nouveaux ; on ne peut en effet utiliser 11 et 12 qui sont déjà des représentations en base 10. On se servira, par exemple, des symboles 0, 1, 2, 3, ..., 9, $\perp$, $\top$.

Les nombres représentés en base dix par 12, 13, 23, 24, 115, 250 seront représentés dans un tel système par :

10, 11, 1$\top$, 20, 97, 18$\perp$.

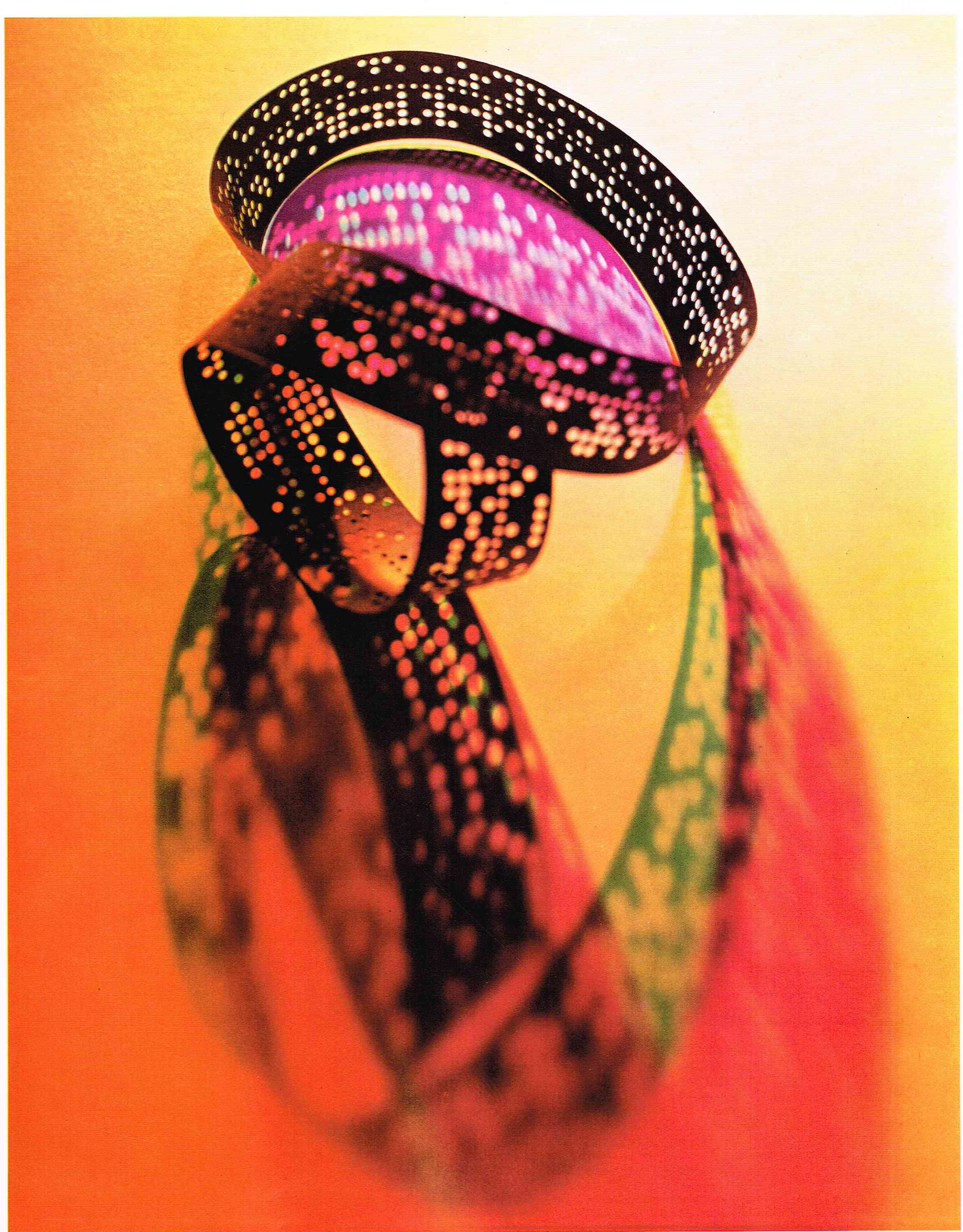
Les mêmes nombres dans le système binaire (base 2, les seuls chiffres utilisés sont 0 et 1) seront représentés par 1 100, 1 101, 10 111, 11 000, 1 110 011, 11 111 010.

Pour déterminer la représentation de 13 (base 10) en base 2, on commence par chercher la plus grande puissance de 2 qui divise 13, ici $13 = 2^3 + 5$, puis la plus grande puissance de 2 qui divise 5 :

$$13 = 2^3 + 2^2 + 1 = 1 \cdot 2^3 + 1 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0$$

d'où la représentation 13 (base 10) $\rightarrow$ 1 101 (base 2).

Les règles d'addition et de multiplication sont bien connues dans le système décimal, mais il n'y a pas de différence fondamentale entre le choix de 10 comme base ou d'un entier quelconque, ces règles sont analogues en base b à ce qu'elles étaient en base 10. On montrera sur un exemple comment se calcule le produit de deux nombres a et b représentés en base 10 par $a = 115$, $b = 24$, en base 12 par $a = 97$, $b = 20$, et en base 2 par $a = 1\,110\,011$, $b = 11\,000$. Alors :


Yan - Rapho

base 10 (A)

	115
×	24
	460
	230
	2760

base 12 (B)

	97
×	20
	00
	172
	1720

base 2 (C)

	1 110 011
	11 000
	0000000
	0000000
	0000000
	1110011
	1110011
	101011001000

Vérification en base 10 de l'opération (B) :

$$2\,760 = 1 \cdot 12^3 + 7 \cdot 12^2 + 2 \cdot 12$$

Vérification en base 10 de l'opération (C) :

$$2\,760 = 2^{11} + 2^9 + 2^7 + 2^6 + 2^3.$$

Les rationnels

A côté des nombres entiers positifs on peut représenter par une suite finie de chiffres inférieurs à la base fixée, et à l'aide de la virgule, tout nombre rationnel du type suivant :

$$(\mathbf{r})\ a = q_m b^m + q_{m-1} b^{m-1} + \ldots + q_1 b + q_0 + q_{-1} b^{-1} + \ldots + q_{-n} b^{-n}$$

La représentation de a dans le système sera :

$$q_m\, q_{m-1} \ldots q_1\, q_0,\, q_{-1}\, q_{-2} \ldots q_{-n}$$

A partir d'une représentation en base b d'un nombre rationnel de ce type, on peut en obtenir une autre en ajoutant après q_{-n} un nombre quelconque de zéros. On confondra ces représentations, et il y aura toujours à cette identification près correspondance biunivoque entre un nombre et sa représentation.

Tous les nombres rationnels n'admettent pas une représentation de ce type, ni *a fortiori* tous les nombres réels. Pour obtenir un développement en base b d'un nombre réel quelconque, on procédera par approximations successives par des rationnels du type cité. Soit a un nombre réel positif, $q_m\, q_{m-1} \ldots q_0$ le plus grand entier inférieur ou égal à a, $q_m\, q_{m-1} \ldots q_0,\, q_{-1}$ le plus grand rationnel du type (**r**) avec un chiffre après la virgule, inférieur ou égal à a, $q_m\, q_{m-1} \ldots q_0,\, q_{-1}\, q_{-2}$ le plus grand rationnel de type (**r**) avec deux chiffres après la virgule inférieur ou égal à a, etc. On engendre ainsi une suite avec une infinité de chiffres après la virgule

$$q_m\, q_{m-1} \ldots q_1\, q_0,\, q_{-1}\, q_{-2} \ldots q_{-n}$$

Une telle suite va représenter le nombre a ; quelle que soit la précision imposée, on pourra approcher le nombre réel a par un tel développement (il suffit de choisir n suffisamment grand).

Deux nombres réels distincts ont deux représentations distinctes. Tout développement $q_m\, q_{m-1} \ldots q_0,\, q_{-1} \ldots q_{-n} \ldots$ représente un nombre réel positif bien déterminé, à savoir le nombre qui sépare la classe des nombres réels inférieurs ou égaux au nombre rationnel $q_m \ldots q_0,\, q_{-1} \ldots q_{-n}$ pour tout entier positif n, de la classe des nombres réels supérieurs ou égaux à ce nombre rationnel.

La représentation d'un nombre réel négatif s'obtient en faisant précéder du signe — le développement de sa valeur absolue.

BIBLIOGRAPHIE

BOREVITCH Z.-I. et CHAPAREVITCH I.-R., *Théorie des nombres,* trad. franç., Paris, Gauthier-Villars, 1967. - BOURBAKI, *Éléments d'histoire des mathématiques,* Paris, Hermann, 1960. - CHEVALLEY C., *Encyclopédie française*. - GODEMENT R., *Cours d'algèbre,* Paris, Hermann, 1966. - ITARD J., *les Nombres premiers,* Paris, Presses universitaires de France, coll. « Que sais-je ? », n° 571, 1969 ; *Arithmétique et Théorie des nombres,* Paris, Presses universitaires de France, coll. « Que sais-je ? », n° 1093, 2e édition 1967. - LE LIONNAIS, *les Grands Courants de la pensée mathématique,* Paris, Blanchard, 1962. - SAMUEL, *Théorie algébrique des nombres,* Paris, Hermann, 1967.

STRUCTURES ALGÉBRIQUES

La notion de *structure algébrique* s'est dégagée progressivement au cours du XIXe siècle ; c'est l'aboutissement d'un processus d'axiomatisation de la recherche qui a conduit les mathématiciens à la création d'êtres mathématiques nouveaux rassemblant les propriétés communes à divers ensembles concrets étudiés jusqu'alors de façon spécifique.

L'étude des structures algébriques constitue l'objet de ce qu'on appelle l'algèbre moderne. Mais les origines de l'algèbre sont beaucoup plus anciennes puisque c'est aux Babyloniens et aux Égyptiens que l'on doit les premières règles de calcul sur les entiers naturels et les nombres rationnels positifs.

Puis les Grecs, avec Euclide et plus tard Diophante, développèrent les règles du calcul algébrique abstrait, mais le champ de leurs recherches, encore très réduit, ne se prêtait pas à l'axiomatisation, le zéro et les nombres négatifs n'apparaissant qu'au haut Moyen Age chez les mathématiciens hindous.

Ce sont les algébristes italiens du XVIe siècle qui introduiront les nombres complexes pour la résolution des équations du 3e et du 4e degré.

D'autre part, la notation algébrique prend la forme que nous lui connaissons aujourd'hui à la fin du XVIe siècle avec Viète, qui, le premier, introduisit des lettres dans les équations algébriques, et surtout avec Descartes.

La fin du XVIIe siècle et le XVIIIe siècle produiront peu de nouvelles découvertes en algèbre, et il faudra attendre les années 1800 pour que les recherches dans ce domaine prennent un nouvel essor (citons en particulier Gauss et Cauchy).

C'est l'œuvre d'Abel et surtout de Galois qui marque le passage de l'algèbre classique à l'algèbre moderne en

3.14159	26535	89793	23846	26433	83279	50288	41971	69399	37510
58209	74944	59230	78164	06286	20899	86280	34825	34211	70679
82148	08651	32823	06647	09384	46095	50582	23172	53594	08128
48111	74502	84102	70193	85211	05559	64462	29489	54930	38196
44288	10975	66593	34461	28475	64823	37867	83165	27120	19091
45648	56692	34603	48610	45432	66482	13393	60726	02491	41273
72458	70066	06315	58817	48815	20920	96282	92540	91715	36436
78925	90360	01133	05305	48820	46652	13841	46951	94151	16094
33057	27036	57595	91953	09218	61173	81932	61179	31051	18548
07446	23799	62749	56735	18857	52724	89122	79381	83011	94912
98336	73362	44065	66430	86021	39494	63952	24737	19070	21798
60943	70277	05392	17176	29317	67523	84674	81846	76694	05132
00056	81271	45263	56082	77857	71342	75778	96091	73637	17872
14684	40901	22495	34301	46549	58537	10507	92279	68925	89235
42019	95611	21290	21960	86403	44181	59813	62977	47713	09960
51870	72113	49999	99837	29780	49951	05973	17328	16096	31859
50244	59455	34690	83026	42522	30825	33446	85035	26193	11881
71010	00313	78387	52886	58753	32083	81420	61717	76691	47303
59825	34904	28755	46873	11595	62863	88235	37875	93751	95778
18577	80532	17122	68066	13001	92787	66111	95909	21642	01989
38095	25720	10654	85863	27886	59361	53381	82796	82303	01952
03530	18529	68995	77362	25994	13891	24972	17752	83479	13151
55748	57242	45415	06959	50829	53311	68617	27855	88907	50983
81754	63746	49393	19255	06040	09277	01671	13900	98488	24012
85836	16035	63707	66010	47101	81942	95559	61989	46767	83744
94482	55379	77472	68471	04047	53464	62080	46684	25906	94912
93313	67702	89891	52104	75216	20569	66024	05803	81501	93511
25338	24300	35587	64024	74964	73263	91419	92726	04269	92279
67823	54781	63600	93417	21641	21992	45863	15030	28618	29745
55706	74983	85054	94588	58692	69956	90927	21079	75093	02955
32116	53449	87202	75596	02364	80665	49911	98818	34797	75356
63698	07426	54252	78625	51818	41757	46728	90977	77279	38000
81647	06001	61452	49192	17321	72147	72350	14144	19735	68548
16136	11573	52552	13347	57418	49468	43852	33239	07394	14333
45477	62416	86251	89835	69485	56209	92192	22184	27255	02542
56887	67179	04946	01653	46680	49886	27232	79178	60857	84383
82796	79766	81454	10095	38837	86360	95068	00642	25125	20511
73929	84896	08412	84886	26945	60424	19652	85022	21066	11863
06744	27862	20391	94945	04712	37137	86960	95636	43719	17287
46776	46575	73962	41389	08658	32645	99581	33904	78027	59009
94657	64078	95126	94683	98352	59570	98258	22620	52248	94077
26719	47826	84826	01476	99090	26401	36394	43745	53050	68203
49625	24517	49399	65143	14298	09190	65925	09372	21696	46151
57098	58387	41059	78859	59772	97549	89301	61753	92846	81382
68683	86894	27741	55991	85592	52459	53959	43104	99725	24680
84598	72736	44695	84865	38367	36222	62609	91246	08051	24388
43904	51244	13654	97627	80797	71569	14359	97700	12961	60894
41694	86855	58484	06353	42207	22258	28488	64815	84560	28506
01684	27394	52267	46767	88952	52138	52254	99546	66727	82398
64565	96116	35488	62305	77456	49803	55936	34568	17432	41125
15076	06947	94510	96596	09402	52288	79710	89314	56691	36867
22874	89405	60101	50330	86179	28680	92087	47609	17824	93858
90097	14909	67598	52613	65549	78189	31297	84821	68299	89487
22658	80485	75640	14270	47755	51323	79641	45152	37462	34364
54285	84447	95265	86782	10511	41354	73573	95231	13427	16610
21359	69536	23144	29524	84937	18711	01457	65403	59027	99344
03742	00731	05785	39062	19838	74478	08478	48968	33214	45713
86875	19435	06430	21845	31910	48481	00537	06146	80674	91927
81911	97939	95206	14196	63428	75444	06437	45123	71819	21799
98391	01591	95618	14675	14269	12397	48940	90718	64942	31961
56794	52080	95146	55022	52316	03881	93014	20937	62137	85595
66389	37787	08303	90697	92077	34672	21825	6259		

Palais de la Découverte, Paris

▶ ***Représentation décimale du nombre π.***

reliant l'étude des équations algébriques à celle des groupes de permutation qui lui sont liés. Diverses notions abstraites commencent à apparaître : celle de groupe, qui domine les premières recherches; celle de loi de composition dégagée par les mathématiciens anglais entre 1830 et 1850; puis les études sur les nombres algébriques, qui sont à l'origine de recherches sur les corps et les anneaux (Dedekind, Hilbert). La synthèse de ces divers courants par l'usage de la méthode axiomatique, qui marque le début de l'algèbre moderne, apparaît pour la première fois dans le livre de Steinitz : *Algebraische Theorie der Körper* (1910); puis l'étude des structures se poursuit et s'enrichit grâce à Artin, Nœther et aux mathématiciens de leur école; le livre de Van der Waerden, *Moderne Algebra* (1930), premier exposé d'ensemble de l'algèbre moderne, a été le point de départ de toutes les recherches d'algèbre contemporaines.

L'objet de ce chapitre est l'étude de certaines structures algébriques fondamentales comme celles des groupes, des anneaux et des corps, ainsi que des notions de base qui s'y rattachent; pour des développements ultérieurs, on pourra se reporter au chapitre *Algèbre linéaire*.

Lois de composition et catégories

Lois de composition interne

Si l'on se fixe un ensemble E, on appelle *loi de composition interne dans E* une application *f* d'une partie A de l'ensemble produit E × E (voir chapitre *Les ensembles*) dans E. Au couple (x, y) de A, la loi de composition interne fait correspondre un élément unique *z* de E :

$$(x, y) \rightarrow z = f(x, y)$$

Si A = E × E, la loi est *partout définie sur E*, c'est alors une *loi interne sur E; z* s'appelle le *composé* de *x* et de *y* pour cette loi, et on le note avec des symboles divers :

$$x + y; \quad x \cdot y \quad \text{ou} \quad xy; \quad x \perp y; \quad x \top y$$

Une loi notée par le signe + s'appelle le plus souvent *addition* (le composé $x + y$ est alors la *somme* de *x* et de *y*); une loi notée par le signe · ou sans signe sera appelée *multiplication* (le composé $x \cdot y$ ou xy est alors le *produit* de *x* et de *y*); une loi quelconque sera généralement notée $x \perp y$ ou $x \top y$.

Citons comme exemples de lois de composition interne partout définies l'addition, la multiplication, l'exponentiation dans l'ensemble ℕ des entiers naturels; mais la soustraction $x - y$ dans ℕ n'est définie que pour les couples (x, y) tels que $x \geqslant y$; de même, la division x/y dans ℕ n'est définie que pour les couples (x, y) tels que *y* soit différent de zéro et que *x* soit multiple de *y*.

Dans ce qui suit, nous ne parlerons, sauf précision contraire, que de lois partout définies.

Associativité

Soit E un ensemble muni d'une loi de composition interne notée ⊤. Étant donné *x*, *y* et *z*, trois éléments de E, on peut calculer les composés $(x \top y) \top z$ et $x \top (y \top z)$. La loi interne ⊤ sera *associative* si, quels que soient les éléments *x*, *y*, *z* de E, on a l'égalité :

$$(x \top y) \top z = x \top (y \top z)$$

le composé est alors noté $x \top y \top z$.

Plus généralement, on peut alors définir le composé de $x_1, x_2, \ldots x_n$ dans cet ordre :

$$x_1 \top x_2 \top \ldots \top x_n.$$

L'addition, la multiplication des nombres de ℕ, ℤ, ℚ, ℝ, ℂ (voir *Théorie des ensembles*) sont associatives; en revanche, l'exponentiation des entiers naturels ne l'est pas : en effet $(2^1)^2 \neq 2^{(1^2)}$.

Commutativité

La loi ⊥ sera *commutative* si, pour tous les éléments *x* et *y* de E, on a l'égalité :

$$x \top y = y \top x$$

Si cette propriété est vérifiée pour un couple (x, y) d'éléments de E, on dit alors que *x* et *y* sont *permutables*.

L'addition et la multiplication des nombres de ℕ, ℤ, ℚ, ℝ, ℂ sont commutatives, mais l'exponentiation des entiers naturels ne l'est pas. Si la loi ⊤ est à la fois associative et commutative, le composé de *n* éléments $x_1, x_2, \ldots, x_n$ de E, dans n'importe quel ordre cette fois, sera défini de manière unique :

$$x_1 \top x_2 \top \ldots \top x_n = x_2 \top x_1 \top \ldots \top x_n = \ldots = x_n \top x_1 \top x_2 \top \ldots \top x_{n-1}$$

La notation additive sera généralement réservée à des lois commutatives.

Élément neutre

Un élément *e* de E sera appelé *élément neutre* pour la loi ⊤ si, pour tout élément *x* de E, on a les égalités :

$$x \top e = e \top x = x.$$

Il existe, comme on le voit facilement d'après la définition, au plus un élément neutre *e* pour une loi donnée; *e* est alors permutable avec tout élément de E; l'élément neutre d'une loi additive (notée par le signe +) est généralement noté 0, celui d'une loi multiplicative (notée par · ou sans signe) sera en général noté 1 (par analogie avec les éléments neutres respectivement pour l'addition et la multiplication dans ℕ, ℤ, ℚ, ℝ, ℂ). Remarquons que l'exponentiation dans ℕ n'admet pas d'élément neutre, non plus que la multiplication dans l'ensemble des nombres pairs.

Éléments réguliers

Lorsque $x = x'$, on a évidemment, pour tout élément *a* de E : $a \top x = a \top x'$ et $x \top a = x' \top a$.

Réciproquement, si :

$$a \top x = a \top x' \quad \text{entraîne} \quad x = x'$$



▲ ***Le mathématicien français* Évariste Galois *(1811-1832). Dans une note rédigée quelques heures avant de mourir dans un duel, Galois résume sa* théorie des équations algébriques.**

⊤ ou ⊥ ou ∗	**: désigne en général une loi de composition dans un ensemble.**
⊤	**se lit souvent « truc ».**
⊥	**se lit souvent « antitruc ».**
∗	**se lit « étoile ».**

quel que soit x appartenant à E, on dira que a est *régulier à gauche* pour la loi $\top$; on définit de même un élément *régulier à droite* (on dit qu'on peut simplifier par a ces égalités).

Si a est régulier à gauche et à droite, il est *régulier* pour la loi $\top$. S'il existe un élément neutre e de la loi $\top$, e est régulier. Tout nombre est régulier pour l'addition dans $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ mais 0 n'est pas régulier pour la multiplication dans ces ensembles; pour l'exponentiation dans $\mathbb{N}$, tout entier naturel autre que 0 et 1 est régulier.

Éléments symétriques

Si $\top$ admet un élément neutre e, on dira qu'un élément x' de E est *symétrique* d'un élément x de E si :

$$x \top x' = x' \top x = e.$$

On dit qu'un élément x est *symétrisable* s'il existe un élément symétrique de x. L'élément neutre est toujours son propre symétrique, c'est d'ailleurs le seul élément symétrisable de l'addition et de la multiplication dans $\mathbb{N}$. Dans l'addition des nombres réels ou complexes, tout élément est symétrisable. Dans la multiplication de $\mathbb{Q}$, $\mathbb{R}$ ou $\mathbb{C}$, tout élément sauf 0 est symétrisable. L'élément symétrique de x pour une loi additive sera noté $-x$ et appelé *opposé* de x, il sera noté x^{-1} (ou $1/x$) pour une loi multiplicative et appelé *inverse* de x.

Ajoutons quelques remarques :

— le composé de x et de x lui-même sera noté additivement :

$$x + x \quad \text{ou} \quad 2x$$

— et multiplicativement :

$$xx \quad \text{ou} \quad x^2$$

— plus généralement, on posera :

$$\underbrace{x + x + \ldots + x}_{n \text{ fois}} = nx \quad \text{et} \quad \underbrace{x \cdot x \cdot \ldots \cdot x}_{n \text{ fois}} = x^n \ (n \in \mathbb{N}).$$

$0x$ désignera l'élément neutre d'une loi additive, et x^0 celui d'une loi multiplicative.

Si A est une partie d'un ensemble E, muni d'une loi de composition interne notée $\top$ telle que :

$$x \in A \quad \text{et} \quad y \in A \text{ entraîne } x \top y \in A$$

on dit que A est *stable* pour $\top$ ou encore que A est une *partie stable* de E pour cette loi. La restriction de la loi $\top$ à une partie stable de A de E est appelée loi *induite par* $\top$ sur A. (*Exemple :* la partie de $\mathbb{N}$ formée des nombres pairs est stable pour l'addition et la multiplication.)

Lois de composition externes et relations entre lois de composition

Lois de composition externes

On appelle loi de composition externe entre éléments d'un ensemble Ω (dit *ensemble des opérateurs* de la loi) et éléments d'un ensemble E une application f de $\Omega \times E$ dans E : à tout couple formé d'un élément α de Ω et d'un élément x de E, on fait correspondre un élément y de E :

$y = f(\alpha, x)$ (noté généralement $y = \alpha x$ ou $y = x \alpha$)

Les éléments de Ω (que nous désignerons par des lettres grecques) sont appelés les *opérateurs* de la loi. Étant donné une loi interne associative sur un ensemble E, notée multiplicativement, l'application qui au couple (n, x) (où n est un entier naturel différent de 0 et x un élément de E) fait correspondre nx ou x^n est une loi de composition externe entre éléments de $\mathbb{N}^*$ et éléments de E. On peut prendre également pour E l'ensemble des vecteurs de $\mathbb{R}^2$ et pour ensemble d'opérateurs l'ensemble $\mathbb{R}$: au couple $(\alpha, \vec{V})$ on fait correspondre le vecteur $\alpha \vec{V}$, multiplication d'un vecteur par un scalaire.

On peut encore prendre pour E l'ensemble des fonctions dérivables de n variables réelles $x_1, \ldots, x_n$ et comme opérateurs les dérivations partielles par rapport aux variables : $\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_n}$; la loi externe fera alors correspondre au couple $\left(\frac{\partial}{\partial x_i}, f\right)$, où $f \in E$, l'application $\frac{\partial f}{\partial x_i}$.

Relations entre lois de composition

Si un ensemble d'opérateurs sur E, Ω, est muni d'une loi interne associative (notée multiplicativement) telle qu'on ait, pour tous les éléments α, β de Ω et x de E, l'égalité :

$$\alpha(\beta x) = (\alpha \beta) x$$

on dira que la loi externe est *associative* par rapport à la loi interne de Ω.

Si c'est l'ensemble E qui est muni d'une loi interne commutative (notée additivement) telle que, pour tous éléments x, y de E et tout α de Ω, on ait l'égalité :

$$\alpha(x + y) = \alpha x + \alpha y$$

on dira que la loi interne est *distributive* par rapport à la loi interne de E.

Si dans E sont définies deux lois de composition, la première associative et commutative (notée additivement), la seconde associative (notée multiplicativement), on dira que la loi multiplicative est *distributive à gauche* par rapport à la loi additive si :

$$x(y + z) = xy + xz.$$

On définit de même la *distributivité à droite* par :

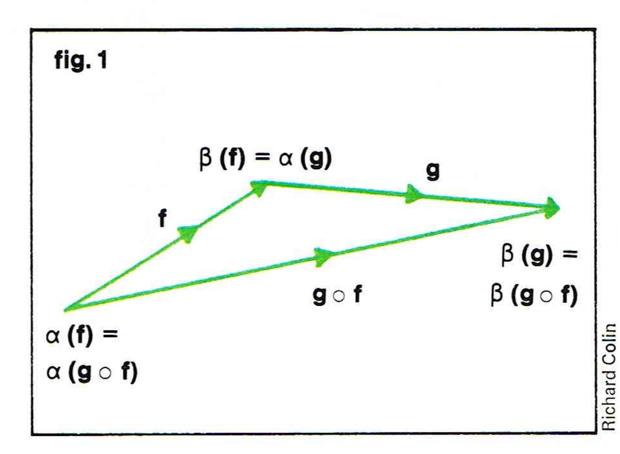
$$(y + z)x = yx + zx$$

Si la loi multiplicative est distributive à gauche et à droite par rapport à la loi additive, elle est *distributive* par rapport à celle-ci. Étant donné une loi multiplicative associative sur un ensemble E, la loi externe $(n, x) \to x^n$ est associative par rapport à la multiplication dans $\mathbb{N}^*$ puisque $(x^m)^n = x^{mn}$; si la loi sur E est également commutative, alors $(x, y)^m = x^m y^m$ et la loi externe est distributive par rapport à la loi interne de E ; la multiplication des nombres appartenant à $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ est distributive par rapport à l'addition de ces nombres.

Nous avons, jusqu'à maintenant, considéré un ensemble E sur lequel étaient définies une (ou plusieurs) loi de composition interne et une (ou plusieurs) loi de composition externe avec un ensemble d'opérateurs Ω ; ces lois et les conditions auxquelles elles sont assujetties (associativité, élément neutre, etc.) déterminent ce qu'on appelle une *structure algébrique* sur E, et la suite de ce chapitre sera consacrée à l'étude de structures algébriques particulières. Nous allons tout d'abord, avant de procéder à une étude plus détaillée des groupes, des anneaux et des corps, donner quelques éléments sur les *catégories*.

Catégories

Un ensemble C muni d'une loi de composition interne (en général non partout définie) associative sera appelé une *catégorie* si les propriétés suivantes sont vérifiées (voir fig. 1) :



▶ **Figure 1 : composition *dans une catégorie* C.**

— il existe deux applications α et β de C dans une partie C_0 de C telles que la restriction de α et β à C_0 soit l'identité (si $f \in C_0$, alors $\alpha(f) = \beta(f) = f$) ;

— le composé $g \cdot f$ de deux éléments de C existe si, et seulement si, $\alpha(g) = \beta(f)$; on a alors :

$$\alpha(g \circ f) = \alpha(f) \quad \text{et} \quad \beta(g \circ f) = \beta(g)$$

— pour tout élément f de C, on a :

$$f \circ \alpha(f) = f = \beta(f) \circ f.$$

Un élément f de C est appelé *sommet* s'il appartient à C_0, *flèche de source* $\alpha(f)$ *et de but* $\beta(f)$ dans le cas contraire. Les applications d'un ensemble dans un autre,

par exemple, forment une catégorie pour la loi de composition habituelle des applications (voir chapitre *Les ensembles*) : une flèche f sera une application d'un ensemble E dans un ensemble E′, sa source $\alpha(f) = id_E$ (application identique de E), son but $\beta(f) = id_{E'}$ (ces applications ne forment pas à proprement parler un ensemble, nous utiliserons néanmoins, par souci de simplification, le terme de « catégorie des applications »).

Foncteur

Un *foncteur* d'une catégorie C dans une catégorie C′ est une application F de l'ensemble C dans l'ensemble C′ telle que :

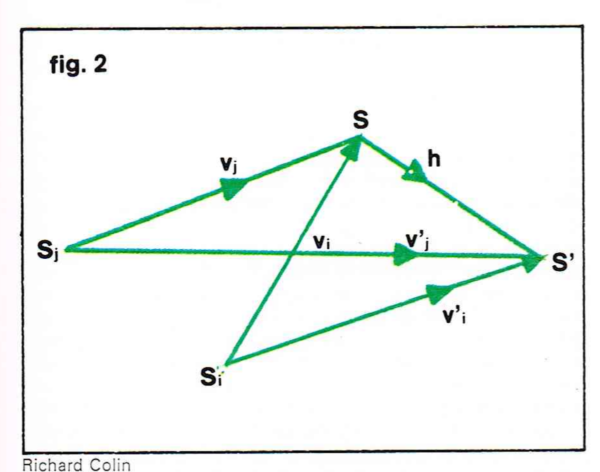
— $F[\alpha(f)] = \alpha[F(f)]$ et $F[\beta(f)] = \beta[F(f)]$
— si $g \cdot f$ est défini dans C, alors
$F(g \cdot f) = F(g) \cdot F(f)$.

On appelle parfois un tel foncteur *foncteur covariant;* un foncteur contravariant est alors une application qui « renverse le sens » des flèches ($F(g \circ f) = F(f) \circ F(g)$).

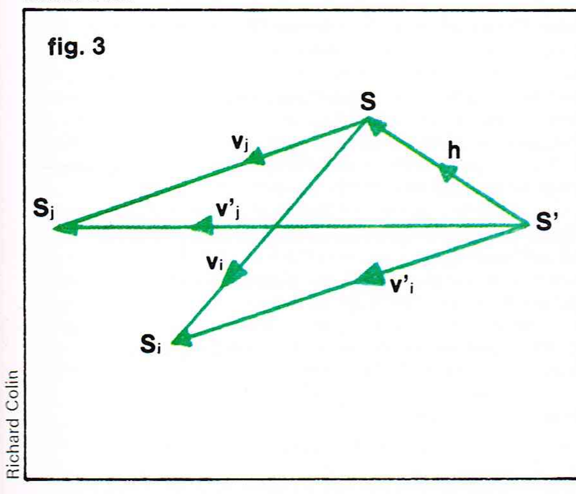
Si l'on considère par exemple la catégorie des homomorphismes de groupes (voir plus loin), l'application qui à un homomorphisme d'un groupe G dans un groupe G′ fait correspondre l'application sous-jacente de l'ensemble G dans l'ensemble G′ est un foncteur de la catégorie des homomorphismes de groupes dans la catégorie des applications; un tel foncteur est appelé *foncteur d'oubli.*

Produits et sommes

Partant d'une catégorie C, considérons une famille $(S_i)_{i \in I}$ de sommets de C_0, on dira qu'un élément S de C_0 est une *somme* (respectivement un *produit*) des S_i si les conditions suivantes sont vérifiées (voir fig. 2 et 3) :



Richard Colin



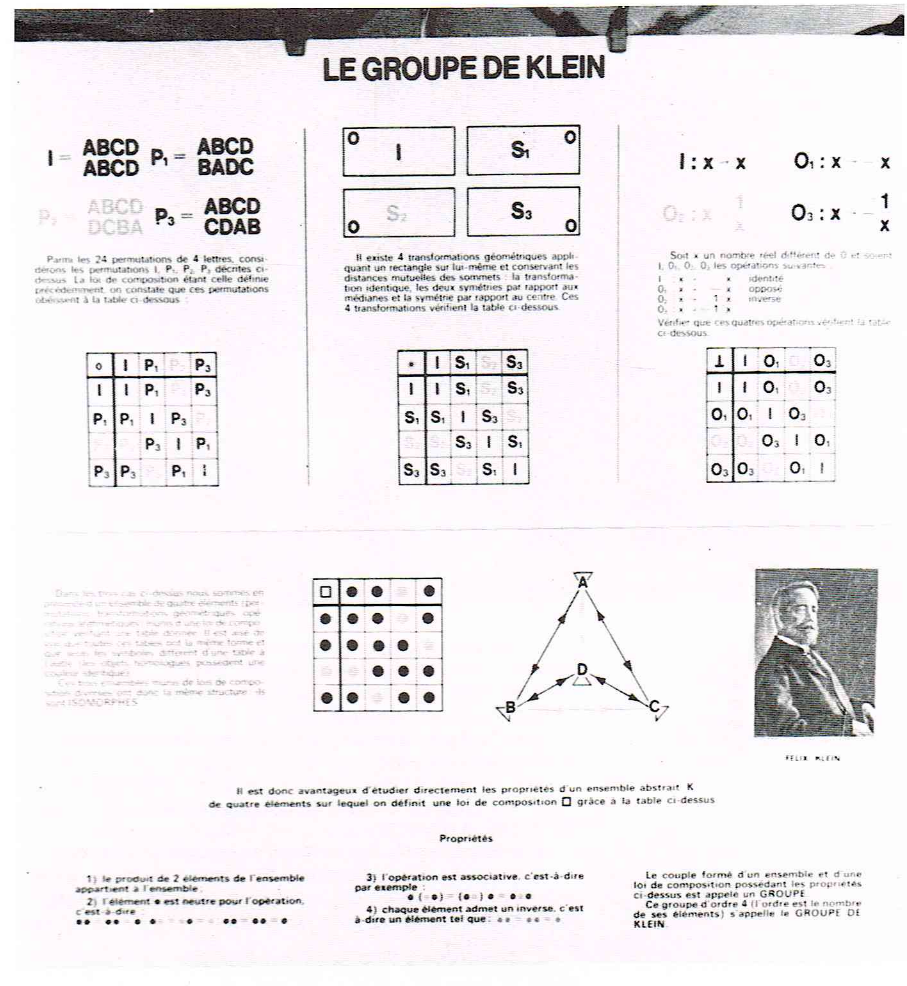
Richard Colin

— pour tout $i \in I$, il existe une flèche v_i de C de source S_i, de but S (respectivement de source S, de but S_i) ;

— si S′ est un autre sommet tel que, pour tout $i \in I$, on ait une flèche v'_i de C de source S_i, de but S′ (respectivement de source S′, de but S_i), alors il existe une flèche unique h de source S et de but S′ (respectivement de source S′, de but S) telle qu'on ait l'égalité :

$h \cdot v_i = v'_i$ (respectivement $v_i \cdot h = v'_i$) pour tout $i \in I$.

Nous verrons par la suite des exemples de produits et de sommes dans des catégories ; notons simplement que le produit cartésien de n ensembles : $A_1 \times A_2 \times \ldots \times A_n$ (voir chapitre *Les ensembles*) est un produit de la famille (A_i) dans la catégorie des applications (l'application identique de chacun des A_i est identifiée à l'ensemble lui-même).



Palais de la Découverte, Paris

▲ *On doit à* Félix Klein, ***mathématicien allemand (1849-1925), d'importants travaux, notamment sur l'application de la*** **théorie des groupes** ***à la géométrie.***

◀ ***En haut,*** **figure 2 :** ***somme S de la famille*** $(S_i)_{i \in I}$ ***où I est un ensemble d'indices quelconque .***
En bas, **figure 3 :** ***produit S de la famille*** $(S_i)_{i \in I}$ ***où I est un ensemble d'indices quelconque.***

Groupes

Groupes et sous-groupes

Groupes

Fixons-nous un ensemble G, sur lequel existe une loi de composition interne partout définie ; on dira que cette loi (que nous noterons multiplicativement) détermine sur G une *structure de groupe* (ou encore que G est un groupe) si elle possède les propriétés suivantes :

— elle est associative : $(xy)z = x(yz)$ (pour tous éléments x, y, z de G) ;

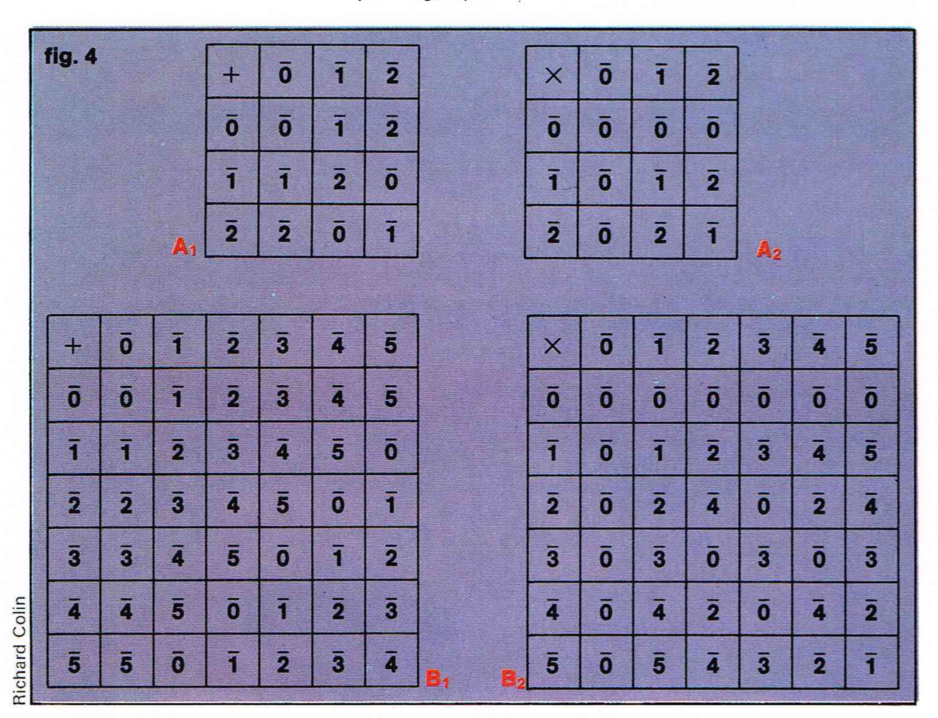
— elle possède un élément neutre unique e : $ex = xe = x$ (pour tout $x \in G$) ;

— chaque élément x de G admet un symétrique unique x^{-1} : $x^{-1}x = xx^{-1} = e$.

Si de plus la loi est commutative ($xy = yx$), le groupe G est alors *commutatif* ou *abélien* (dans ce cas, la loi sera souvent notée additivement). Un groupe G est *fini* s'il comporte un nombre fini d'éléments n, n est alors l'*ordre* du groupe ; il est *infini* dans le cas contraire.

Comme exemples de groupes, citons les ensembles $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ pour l'addition (l'ensemble $\mathbb{N}$ des entiers naturels n'est pas un groupe pour l'addition puisque aucun élément, sauf 0, n'admet de symétrique) ; si on considère la multiplication sur $\mathbb{R}$, elle vérifie les deux premières propriétés, mais 0 n'admet pas de symétrique : c'est l'ensemble $\mathbb{R}$ moins zéro qui est un groupe pour la multiplication. Tous ces groupes sont des groupes abéliens infinis. Le groupe $\mathbb{Z}/n\mathbb{Z}$ des entiers *modulo n*, ensemble quotient de $\mathbb{Z}$ par la relation d'équivalence $\mathcal{R} : x\mathcal{R}y$ si, et seulement si $x - y = nz$ $(z \in \mathbb{Z})$ (voir

chapitre *Les ensembles*), dont les éléments sont donc les classes d'équivalence suivant $\mathcal{R}$, est un groupe pour l'addition (voir ci-dessous) ; c'est un groupe fini d'ordre n (voir fig. 4).



A1:

+	$\bar{0}$	$\bar{1}$	$\bar{2}$
$\bar{0}$	$\bar{0}$	$\bar{1}$	$\bar{2}$
$\bar{1}$	$\bar{1}$	$\bar{2}$	$\bar{0}$
$\bar{2}$	$\bar{2}$	$\bar{0}$	$\bar{1}$

A2:

×	$\bar{0}$	$\bar{1}$	$\bar{2}$
$\bar{0}$	$\bar{0}$	$\bar{0}$	$\bar{0}$
$\bar{1}$	$\bar{0}$	$\bar{1}$	$\bar{2}$
$\bar{2}$	$\bar{0}$	$\bar{2}$	$\bar{1}$

B1:

+	$\bar{0}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$	$\bar{5}$
$\bar{0}$	$\bar{0}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$	$\bar{5}$
$\bar{1}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$	$\bar{5}$	$\bar{0}$
$\bar{2}$	$\bar{2}$	$\bar{3}$	$\bar{4}$	$\bar{5}$	$\bar{0}$	$\bar{1}$
$\bar{3}$	$\bar{3}$	$\bar{4}$	$\bar{5}$	$\bar{0}$	$\bar{1}$	$\bar{2}$
$\bar{4}$	$\bar{4}$	$\bar{5}$	$\bar{0}$	$\bar{1}$	$\bar{2}$	$\bar{3}$
$\bar{5}$	$\bar{5}$	$\bar{0}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$

B2:

×	$\bar{0}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$	$\bar{5}$
$\bar{0}$	$\bar{0}$	$\bar{0}$	$\bar{0}$	$\bar{0}$	$\bar{0}$	$\bar{0}$
$\bar{1}$	$\bar{0}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$	$\bar{5}$
$\bar{2}$	$\bar{0}$	$\bar{2}$	$\bar{4}$	$\bar{0}$	$\bar{2}$	$\bar{4}$
$\bar{3}$	$\bar{0}$	$\bar{3}$	$\bar{0}$	$\bar{3}$	$\bar{0}$	$\bar{3}$
$\bar{4}$	$\bar{0}$	$\bar{4}$	$\bar{2}$	$\bar{0}$	$\bar{4}$	$\bar{2}$
$\bar{5}$	$\bar{0}$	$\bar{5}$	$\bar{4}$	$\bar{3}$	$\bar{2}$	$\bar{1}$

▲ **Figure 4 ;** A : ***addition (1) et multiplication (2) dans*** $\mathbb{Z}/3\mathbb{Z}$; B : ***addition (1) et multiplication (2) dans*** $\mathbb{Z}/6\mathbb{Z}$.

Si E est un ensemble, considérons l'ensemble des applications biunivoques (ou bijections) de E sur lui-même ; cet ensemble est muni d'une structure de groupe pour la composition des applications ; en effet :

— $f \circ (g \circ h) = (f \circ g) \circ h$ (quelles que soient les applications f, g, h) ;

— l'application identique de E est élément neutre ;

— chaque élément admet un symétrique unique, l'application réciproque.

Le groupe, noté S_E, s'appelle *groupe symétrique* de l'ensemble E. Si l'ensemble E comporte n éléments, le groupe symétrique de E est encore appelé *groupe de substitution*, et on le note alors S_n.

Sous-groupes

Fixons-nous un groupe (multiplicatif) G. Une partie H de G sera un *sous-groupe* de G, si H est lui-même un groupe pour la loi induite par celle de G. On a les résultats suivants : H est un sous-groupe de G si, et seulement si, l'une ou l'autre des conditions suivantes est vérifiée :

(1) H est une partie stable de G (si $x \in$ H et $y \in$ H, alors $xy \in$ H), et le symétrique de tout élément de H appartient encore à H ;

(2) si $x \in$ H et $y \in$ H, alors $xy^{-1} \in$ H.

En effet, si H est un groupe, il est évidemment stable pour la multiplication de G puisque la loi induite sur H par celle de G doit être partout définie ; l'élément neutre e' de H est le même que e, élément neutre de G, puisqu'on a : $e'e' = e'$ et donc $e' = e'(e'e'^{-1}) = e'e'^{-1} = e$; il en résulte que l'élément symétrique (ou inverse) de x dans H est identique à son inverse dans G, ce qui démontre **(1)**, d'où on déduit immédiatement **(2)**. Réciproquement, partant de l'hypothèse **(2)**, on en déduit que l'élément neutre e de G appartient à H (il suffit de prendre $y = x$) d'où, si $y \in$ H, alors $y^{-1} \in$ H (on prend $x = e$), et la stabilité se déduit immédiatement : $x(y^{-1})^{-1} = xy \in$ H ; il en résulte qu'il existe bien sur H une loi de composition interne partout définie, évidemment associative (puisqu'elle possède cette propriété dans G), admettant un élément neutre e, et telle que tout élément de H possède un symétrique dans H : H est un groupe.

Nous avons vu que les ensembles $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ sont des groupes pour l'addition ; il est immédiat que, pour cette même loi, $\mathbb{Z}$ est un sous-groupe de $\mathbb{Q}$ qui est lui-même un sous-groupe de $\mathbb{R}$, lui-même un sous-groupe de $\mathbb{C}$. Un sous-groupe du groupe additif $\mathbb{Z}$ qui n'est pas réduit au seul élément 0 sera formé de tous les entiers de la forme pn (multiples de n) où $p \in \mathbb{Z}$ et n est le plus petit élément positif de ce sous-groupe (un tel ensemble est encore noté $n\mathbb{Z}$).

Étant donné un groupe multiplicatif G et un élément x de G distinct de e, considérons l'ensemble des éléments de G de la forme x^n pour $n \in \mathbb{Z}$. Cet ensemble constitue un sous-groupe H de G dont x est le *générateur*. Ce groupe H, que l'on appelle *groupe cyclique* ou *groupe monogène*, est évidemment abélien.

Il peut être :

soit infini, et c'est alors l'ensemble formé des éléments :

$$\dots x^{-n}, \dots, x^{-1}, x^0 = e, x, \dots x^n, \dots$$

soit fini dans le cas où il existe un entier positif tel que $x^n = e$; si n est le plus petit entier possédant cette propriété, H est alors un groupe fini d'ordre n formé des éléments :

$$e, x, \dots, x^{n-1}.$$

Citons comme exemple de groupe cyclique fini le groupe multiplicatif engendré par -1 : c'est un groupe d'ordre 2 formé des éléments -1 et $+1$.

Groupes quotients

Classes suivant un sous-groupe

Fixons-nous un groupe multiplicatif G et supposons définie une relation d'équivalence $\mathcal{R}$ sur G. On dira que la relation d'équivalence $\mathcal{R}$ est *compatible à gauche* avec la loi du groupe si la propriété suivante est vérifiée :

si x et x', éléments de G, sont tels que $x \mathcal{R} x'$, alors $yx \mathcal{R} yx'$ pour tout $y \in$ G.

On définit de même une relation *compatible à droite* : $x \mathcal{R} x'$ entraîne $xy \mathcal{R} x'y$ (pour tout $y \in$ G).

Si $\mathcal{R}$ est compatible à droite et à gauche avec la loi de G, on dit que $\mathcal{R}$ est *compatible* avec cette loi ; elle vérifie dans ce cas :

si $x \mathcal{R} x'$ et $y \mathcal{R} y'$, alors $xy \mathcal{R} x'y'$.

Supposons maintenant donnée une relation d'équivalence $\mathcal{R}$ sur G compatible à gauche ; si x et y de G sont tels que $x \mathcal{R} y$, on aura aussi $e = x^{-1}x \mathcal{R} x^{-1}y$ et donc $x^{-1}y$ appartient à la classe d'équivalence H de e (H est la partie de G formée des éléments x tels que $x \mathcal{R} e$) ; montrons que H est un sous-groupe de G, c'est-à-dire, comme nous l'avons vu, que si $x \in$ H et $y \in$ H, alors $xy^{-1} \in$ H : $y \in$ H s'écrit encore $y \mathcal{R} e$, d'où $e = y^{-1}y \mathcal{R} y^{-1}$, et donc $xy^{-1} \mathcal{R} x$; par transitivité (voir chapitre *Les ensembles*), on obtient, puisque $x \mathcal{R} e$, $xy^{-1} \mathcal{R} e$ et donc $xy^{-1} \in$ H. Réciproquement, on montre facilement que, si H est un sous-groupe quelconque de G, la relation $x \mathcal{R} y$ si, et seulement si $x^{-1}y \in$ H est une relation d'équivalence compatible à gauche avec la loi du groupe, d'où le résultat.

Toute relation d'équivalence $\mathcal{R}$ compatible à gauche avec la loi d'un groupe G est de la forme $x^{-1}y \in$ H, H étant un sous-groupe quelconque de G ; de même, toute relation d'équivalence $\mathcal{R}$ compatible à droite avec la loi d'un groupe G est de la forme $yx^{-1} \in$ H, H étant un sous-groupe quelconque de G. La classe d'un élément x de G suivant une relation compatible à gauche sera donc formée des éléments xy de G tels que $y \in$ H : on la note xH ; Hx sera la classe de x suivant une relation compatible à droite ; ces classes s'appellent respectivement *classes à gauche* et *classes à droite suivant* (ou *modulo*) H.

Toutes ces classes ont même puissance puisqu'on peut établir une correspondance biunivoque entre xH et yH ; en particulier, si G est d'ordre fini, comme H l'est nécessairement aussi, et que les classes suivant H n'ont pas d'éléments communs, on en déduit la propriété suivante :

dans un groupe fini, l'ordre d'un sous-groupe quelconque est un diviseur de l'ordre du groupe.

Sous-groupes distingués et groupes quotients

Un sous-groupe H de G s'appelle *sous-groupe distingué* (ou *invariant*) de G si l'on a :

$$x\text{H}x^{-1} = \text{H} \quad (\text{pour tout } x \in \text{G}) ;$$

un sous-groupe H sera distingué si, pour tout $y \in$ H et tout $x \in$ G, on a : $xyx^{-1} \in$ H.

Les classes à gauche et à droite suivant H sont alors confondues, puisque $x\text{H}x^{-1} = \text{H}$ entraîne $x\text{H} = \text{H}x$ et la relation d'équivalence $\mathcal{R}$:

$x^{-1}y \in$ H est compatible avec la loi du groupe.

Si l'on considère l'ensemble quotient $\text{G}/\mathcal{R}$ dont les éléments sont les classes d'équivalence de G suivant $\mathcal{R}$, on peut définir sur cet ensemble la loi de composition

suivante ($\bar{x}$ désigne la classe suivant $\mathcal{R}$ d'un élément quelconque x de G) :

$$\overline{xy} = \bar{x}\bar{y}.$$

Cette loi est bien partout définie sur $G/\mathcal{R}$ et vérifie de plus que :

— elle est associative (puisque la loi de G l'est) ;

— elle possède un élément neutre $\bar{e}$;

— tout élément $\bar{x}$ admet un symétrique $\bar{x}^{-1} = \overline{x^{-1}}$.

$G/\mathcal{R}$ est donc un groupe encore noté G/H et appelé *groupe quotient* de G par le sous-groupe distingué H.

Notons que si la loi de composition d'un groupe G est commutative, pour tous les éléments x, y de G on a : $xyx^{-1} = y$; donc tout sous-groupe d'un groupe abélien est distingué ; en particulier, considérons le groupe additif $\mathbb{Z}$: nous avons vu que ses sous-groupes sont de la forme $n\mathbb{Z}$ avec $n \in \mathbb{N}^*$; la relation $x - y \in n\mathbb{Z}$ s'écrit encore :

$$x \equiv y \ (modulo\ n)$$

et se lit x est *congru* à y *modulo n* ; les groupes quotients de $\mathbb{Z}$ par un sous-groupe $n\mathbb{Z}$ seront notés $\mathbb{Z}/n\mathbb{Z}$, et un tel groupe s'appelle *groupe additif des entiers rationnels modulo n ;* c'est un groupe fini d'ordre n (fig. 4 A_1 et B_1).

Homomorphismes de groupes produits

Homomorphismes de groupes

Une application f d'un groupe G dans un groupe G′ (dont les lois sont notées multiplicativement) est un *homomorphisme* de groupes si :

$$f(xy) = f(x) f(y)$$

(pour tout x, tout y appartenant à G) ; dans ce cas, si e désigne l'élément neutre de G, e' celui de G′

— $e' = f(e)$; $f(x^{-1}) = [f(x)]^{-1}$;

— f (G) est un sous-groupe de G′ appelé *image* de f noté Im (f) ;

— l'ensemble $f^{-1}(e')$ est un sous-groupe distingué de G, appelé *noyau* de l'homomorphisme et noté Ker (f) ; rappelons que $f^{-1}(e')$ est formé des éléments x de G tels que $f(x) = e'$.

Si, de plus, l'application f est bijective, f est alors un *isomorphisme,* et l'application réciproque f^{-1} est un isomorphisme de G′ sur G ; les groupes G et G′ sont alors dits *isomorphes :* en particulier, tout groupe cyclique G d'ordre n est isomorphe à $\mathbb{Z}/n\mathbb{Z}$ (si $G = \{e, x, x^2, ..., x^{n-1}\}$, l'application f telle que

$$f(e) = \bar{0}, f(x) = \bar{1}, ..., f(x^{n-1}) = \overline{n-1}$$

est un isomorphisme de G sur $\mathbb{Z}/n\mathbb{Z}$).

Un homomorphisme de G dans lui-même s'appelle un *endomorphisme* de G, et un endomorphisme bijectif s'appelle un *automorphisme* de G : si a est un élément fixé d'un groupe G, l'application f_a de G dans G définie par $f_a(x) = axa^{-1}$ est un automorphisme de G ; l'ensemble de tels automorphismes f_a pour $a \in G$ forme un groupe pour la loi de composition des applications ; ce groupe s'appelle groupe des *automorphismes intérieurs* de G.

Si H est un sous-groupe distingué de G, l'application f de G sur G/H définie par $f(x) = \bar{x}$ est un homomorphisme surjectif que l'on appelle *homomorphisme canonique* de G sur G/H ; si maintenant f est un homomorphisme de G dans un groupe G′, appelons s l'homomorphisme canonique de G sur G/Ker (f) (nous avons vu que Ker (f) est un sous-groupe distingué de G) ; le groupe quotient G/Ker (f) est isomorphe à l'image f (G) par l'application k :

$$k(\bar{x}) = f(x) ;$$

d'autre part, l'application identique : $i(x') = x'$ de f (G) dans G′ est un homomorphisme injectif (canonique). Nous avons alors procédé à ce qu'on appelle la *décomposition canonique* de l'homomorphisme f, c'est-à-dire écrit f comme le composé de trois homomorphismes canoniques : $f = i \circ k \circ s$ où

— s est l'homomorphisme canonique de G sur Ker (f) ;

— k est l'isomorphisme canonique de G/Ker (f) sur f (v) ;

— i est l'homomorphisme canonique de f (G) dans G′.

Groupe produit

Comme le composé de deux homomorphismes est encore un homomorphisme, les homomorphismes de groupes forment une catégorie pour la composition des applications (sous la réserve faite précédemment) ; considérons n sommets (ou objets) de cette catégorie ; $G_1, ..., G_n$; chacun des G_i est un groupe (identifié à son application identique), et l'ensemble produit $G = G_1 \times G_2 \times ... \times G_n$ (encore noté $\prod_{i=1,...,n} G_i$) peut être muni d'une structure de groupe de la façon suivante : considérons sur G la loi de composition donnée par la formule

$$(x_1, ..., x_n)\ (y_1, ..., y_n) = (x_1y_1, ..., x_ny_n)$$

— elle est associative (cela résulte de l'associativité des lois de $G_1, ..., G_n$) ;

— elle possède un élément neutre $e = (e_1, ..., e_n)$ (où e_i désigne l'élément neutre de G_i pour $1 \leqslant i \leqslant n$) ;

— tout $x = (x_1, ..., x_n)$ admet un inverse

$$x^{-1} = (x_1^{-1}, ..., x_n^{-1}).$$

Le groupe G s'appelle *groupe produit* des groupes $G_1, ..., G_n$ (le résultat, ainsi que ce qui va suivre, est encore valable pour une famille quelconque $(G_i)_{i \in I}$ de groupes).

Pour tout i, l'application p_i de G dans G_i définie par $p_i(x_1, ..., x_n) = x_i$ est évidemment un homomorphisme de groupes ; supposons de plus donnés un groupe G′ ainsi que n homomorphismes $p'_1, ..., p'_n$ de G′ dans les groupes G_i, il existe alors un unique homomorphisme h de G′ dans G tel que pour tout $i = 1, ..., n$, on ait $p_i \cdot h = p_i$; c'est l'homomorphisme : $h(x') = (p_1'(x'), ..., p_n'(x'))$ pour tout $x' \in G'$. Le groupe G est donc un produit de $G_1, ..., G_n$ dans la catégorie des homomorphismes de groupes.

Anneaux et corps

Anneaux

Un ensemble A muni de deux lois de composition internes (une addition et une multiplication) est appelé un *anneau* si :

— A est un groupe abélien pour l'addition ;

— la multiplication est associative ;

— la multiplication est distributive par rapport à l'addition.

Les hypothèses sur l'addition dans A s'expriment par les égalités :

$$x + (y + z) = (x + y) + z \ (associativité)$$
$$x + y = y + x \ (commutativité)$$

par l'existence d'un élément neutre noté 0 tel que :

$$x + 0 = x$$

et par l'existence, pour tout x, d'un élément opposé à x, $-x$, tel que :

$$x + (-x) = 0 \quad \text{ou encore} \quad x - x = 0.$$

Les hypothèses sur la multiplication s'expriment par les égalités :

$$x(yz) = (xy)z \quad (associativité) ;$$
$$\left.\begin{array}{l} x(y+z) = xy + xz \\ (y+z)x = yx + zx \end{array}\right\} (distributivité).$$

Si la multiplication est commutative, on dit que l'anneau est *commutatif* ou *abélien.*

Si la multiplication possède un élément neutre, cet élément, qui s'appelle *élément unité* de A, est alors en général noté 1 et A est appelé *anneau unitaire.* Nous avons vu que $\mathbb{Z}$ est un groupe abélien pour l'addition, et que la multiplication sur $\mathbb{Z}$ est associative et distributive par rapport à l'addition ; $\mathbb{Z}$, muni de ces deux lois, est donc un anneau que l'on appelle *anneau des entiers rationnels ;* c'est un anneau commutatif unitaire (d'élément unité + 1) ; de même, comme nous le verrons par la suite, pour tout $n \in \mathbb{N}^*$, $\mathbb{Z}/n\mathbb{Z}$ est un anneau abélien unitaire.

Nous allons voir maintenant quelques propriétés des anneaux :

— $y + (-x)$ noté $y - x$ nous donne la soustraction, opération inverse de l'addition et par rapport à laquelle la multiplication est distributive ;

— le produit d'un élément quelconque de A par 0 est

toujours égal à 0 ($x0 = x(y - y) = xy - xy = 0$), mais inversement, il peut exister des éléments x, y de H différents de 0 tels que $xy = 0$; de tels éléments sont alors appelés *diviseurs* de zéro. Un anneau commutatif, non réduit à 0 et dépourvu de diviseurs de zéro, est appelé *anneau intègre* ou *anneau d'intégrité* : l'anneau $\mathbb{Z}$ est un anneau intègre mais l'anneau $\mathbb{Z}/6\mathbb{Z}$ (voir fig. 4B₂) a des diviseurs de zéro : $\bar{2} \cdot \bar{3} = \bar{0}$, on voit qu'un anneau $\mathbb{Z}/n\mathbb{Z}$ est intègre si, et seulement si n est premier ;

— si A est abélien, la formule du binôme de Newton est vérifiée :

$$(x + y)^n = x^n + C_n x^{n-1} y + ... + C_n x^{n-p} y^p + ... + C_n^{n-1} x y^{n-1} + y^n \quad (n \in \mathbb{N}^*)$$

cette formule étant encore valable dans un anneau non commutatif si x et y sont permutables (c'est-à-dire si $xy = yx$) ;

— s'il existe un entier positif n tel que, pour tout élément x de A, on ait $nx = 0$, on dit que A est de *caractéristique non nulle* ; si n est le plus petit entier vérifiant cette propriété, A est de caractéristique n ; si ce n'est le cas d'aucun entier, l'anneau est de *caractéristique nulle* : $\mathbb{Z}$ est de *caractéristique nulle*, alors que $\mathbb{Z}/n\mathbb{Z}$, pour n premier, est de caractéristique n.

Sous-anneaux

Si une partie B d'un anneau A est elle-même un anneau pour les lois induites sur B par celles de A, on dit que B est un sous-anneau de A ; les sous-anneaux de A sont les parties B de A qui vérifient les deux propriétés :

— B est un sous-groupe du groupe additif de A ;

— B est stable pour la multiplication.

$\mathbb{Z}$ est un sous-anneau de $\mathbb{Q}$ qui est lui-même un sous-anneau de $\mathbb{R}$, lui-même un sous-anneau de $\mathbb{C}$; les sous-groupes $n\mathbb{Z}$ de $\mathbb{Z}$ sont des sous-anneaux de $\mathbb{Z}$ (on peut remarquer que, bien que $\mathbb{Z}$ soit un anneau unitaire, pour $n \geqslant 2$ ces anneaux ne sont pas unitaires ; en revanche, les sous-anneaux d'un anneau commutatif ou intègre sont eux-mêmes respectivement commutatifs ou intègres).

Idéaux

Si B est un sous-anneau de A, la relation $x \mathcal{R} y$ si, et seulement si $x - y \in B$ est une relation d'équivalence compatible avec l'addition mais, en général, pas avec la multiplication ; en effet, si $\mathcal{R}$ est compatible à gauche avec la multiplication, on doit avoir pour tout élément z de A

$$x \mathcal{R} y \text{ entraîne } zx \mathcal{R} zy$$

ou encore

$$x - y \in B \text{ entraîne } zx - zy \in B \text{ (ou } z(x - y) \in B)$$

ce qui n'est généralement pas vrai pour un sous-anneau quelconque ; pour obtenir une relation compatible avec la multiplication, on est donc conduit à considérer des sous-anneaux particuliers.

On appelle *idéal à gauche* de A une partie I de A qui vérifie les deux propriétés suivantes :

(1) — I est un sous-groupe du groupe additif de A ($a \in I$ et $b \in I$ entraîne $a - b \in I$)

(2) — pour tout élément x de A, $xI = I$ (si $x \in A$ et $a \in I$ alors $xa \in I$)

On définit de même un *idéal à droite* par la propriété **(1)** et la propriété :

(2') — pour tout élément x de A, $Ix = I$.

Un idéal à gauche et à droite est appelé *idéal bilatère* de A (remarquons que, si A est commutatif, tout idéal de A est bilatère).

Si A est un anneau quelconque, $\{0\}$ et A sont des idéaux bilatères de A ; un idéal distinct de $\{0\}$ et A s'appelle *idéal propre* de A ; les idéaux de $\mathbb{Z}$ sont les sous-anneaux $n\mathbb{Z}$, car, pour tout élément p de $\mathbb{Z}$, $p(na) = n(pa) \in n\mathbb{Z}$.

Voyons maintenant quelques propriétés des idéaux :

— l'intersection d'une famille d'idéaux à gauche est encore un idéal à gauche ; en particulier, l'intersection de tous les idéaux à gauche qui contiennent une partie donnée H de A est encore un idéal à gauche, appelé idéal à gauche *engendré* par H ; on dit alors que H est un *système de générateurs* de cet idéal (ces propriétés sont encore valables si on remplace partout « idéal à gauche » par « idéal à droite » ou « idéal bilatère ») ;

— si a est un élément fixé de A, aA (formé des éléments ax où $x \in A$) est un idéal à droite, Aa (formé des éléments xa où $x \in A$) est un idéal à gauche de A. Si A est un anneau unitaire, l'idéal à gauche Aa contient a, et il est clair que tout idéal à gauche contenant a contiendra aussi Aa : Aa est l'idéal à gauche engendré par a ; si, de plus, A est commutatif, l'idéal $aA = Aa$ engendré par a se note (a) et s'appelle *idéal principal ;* un anneau intègre unitaire dont tout idéal est principal s'appelle un anneau principal : $\mathbb{Z}$ est un anneau principal ;

— si A est un anneau commutatif, on appellera *idéal maximal* un idéal I qui possède les propriétés suivantes :

1) I est différent de A ;

2) il n'existe pas d'idéal de A, différent de A et de I lui-même, qui contienne I.

Ces propriétés peuvent encore s'exprimer :

— l'idéal I est maximal si I est un élément maximal pour l'inclusion de l'ensemble des idéaux de A distincts de A (voir chapitre *Les ensembles*) ;

— dans $\mathbb{Z}$, l'idéal $6\mathbb{Z}$ n'est pas maximal, car il est contenu dans les idéaux $2\mathbb{Z}$ et $3\mathbb{Z}$, mais $7\mathbb{Z}$ est maximal : un idéal $p\mathbb{Z}$ de $\mathbb{Z}$ sera maximal si, et seulement si p est premier.

Anneaux quotients

Nous avons vu, c'est ce qui nous a amené à la définition d'un idéal, que, si $\mathcal{R}$ est une relation d'équivalence sur un anneau A compatible avec l'addition :

$$x \mathcal{R} y \text{ si, et seulement si, } x - y \in I$$
(où I est un sous-groupe de A),

elle n'est compatible à gauche avec la multiplication que si I est un idéal à gauche de A ; de même, elle n'est compatible à droite avec la multiplication que si I est un idéal à droite de A. Pour que $\mathcal{R}$ soit compatible avec l'addition et la multiplication, il faut et il suffit donc que I soit un idéal bilatère de A. L'ensemble quotient $A/\mathcal{R}$ est alors un anneau si on le munit de l'addition :

$$\bar{x} + \bar{y} = \overline{x + y}$$

pour laquelle, nous l'avons vu, $A/\mathcal{R}$ possède une structure de groupe et de la multiplication :

$$\bar{x}\bar{y} = \overline{xy}$$

qui est bien associative et distributive par rapport à l'addition. On appelle $A/\mathcal{R}$ *anneau quotient* de A par I, et on le note A/I.

Homomorphismes d'anneaux

Une application f d'un anneau A dans un anneau A' est un homomorphisme d'anneaux si :

$$f(x + y) = f(x) + f(y) \quad \text{et} \quad f(xy) = f(x) f(y)$$

pour tous les éléments x, y de A. Dans ce cas, on a aussi :

— $f(0) = 0$; $\quad f(-x) = -f(x)$;

— f (A) est un sous-anneau de A' ;

— l'ensemble $f^{-1}(0)$ est un idéal bilatère de A appelé noyau de l'homomorphisme ;

— en outre, si A et A' sont unitaires d'éléments unités respectifs e et e', $f(e) = e'$.

L'application de A sur A/I, où I est un idéal bilatère de A, définie par : $f(x) = \bar{x}$, est un homomorphisme surjectif, appelé *homomorphisme canonique* de A sur A/I.

Comme dans le cas des homomorphismes de groupes, un homomorphisme bijectif d'un anneau A sur un anneau A' s'appelle un *isomorphisme* de A sur A' (A et A' sont alors *isomorphes*) ; un homomorphisme de A dans lui-même s'appelle un *endomorphisme* de A et un *automorphisme* de A s'il est bijectif.

Si f est un homomorphisme d'un anneau A dans un anneau A', on peut procéder, comme dans le cas des groupes a', à sa décomposition canonique.

Si f est un homomorphisme d'un anneau A dans un anneau A', g un homomorphisme de A' dans un anneau A'', le composé $g \circ f$ est encore un homomorphisme d'anneaux : les homomorphismes d'anneaux forment une catégorie pour la composition des applications ; si $A_1, ..., A_n$ sont n anneaux, considérons l'ensemble produit $A = A_1 \times ... \times A_n$; nous avons vu que A est un groupe pour l'addition :

$$(x_1, ..., x_n) + (y_1, ..., y_n) = (x_1 + y_1, ..., x_n + y_n) ;$$

si, de plus, on considère la multiplication définie par :

$$(x_1, ..., x_n) \cdot (y_1, ..., y_n) = (x_1 y_1, ..., x_n y_n),$$

ces deux lois confèrent à A une structure d'anneau ; on voit facilement que A, *anneau produit* de $A_1, ..., A_n$, est leur produit dans la catégorie des homomorphismes d'anneaux (ces résultats sont également valables pour une famille quelconque $(A_i)_{i \in I}$ d'anneaux).

Corps

Certains éléments d'un anneau peuvent admettre un symétrique pour la multiplication : c'est le cas, dans $\mathbb{Z}$, de -1 et de 1 ; nous allons voir qu'un corps est un anneau caractérisé par la propriété que tous ses éléments, sauf le zéro de l'addition, possèdent un inverse pour la multiplication.

Un ensemble K, muni d'une addition et d'une multiplication, est un corps s'il possède les deux propriétés :

— K est un anneau pour ces deux lois ;

— les éléments de K distincts du zéro de l'addition forment un groupe pour la multiplication.

L'ensemble des éléments distincts de zéro d'un corps K se notera généralement K^* : c'est le *groupe multiplicatif* du corps ; on peut remarquer qu'un corps possède au moins deux éléments : les éléments unités de chacune des deux lois.

Un corps est dit *commutatif* si son groupe multiplicatif est abélien.

Un corps K est de *caractéristique n* si l'anneau K est de caractéristique n.

L'anneau $\mathbb{Z}$ n'est pas un corps puisque ses éléments n'admettent en général pas d'inverse pour la multiplication ; en revanche, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ sont des corps commutatifs.

Voyons maintenant quelques propriétés des corps :

— dans un corps, un produit est nul si, et seulement si l'un des facteurs au moins est nul ; ce qui revient à dire qu'un corps ne peut avoir de diviseurs de zéro ; en effet, si :

$$ab = 0 \text{ (avec } a \neq 0)$$

comme a admet un inverse a^{-1} pour la multiplication,

$$a^{-1}ab = b = 0.$$

Nous avons vu que, si n n'est pas premier, les anneaux $\mathbb{Z}/n\mathbb{Z}$ possèdent des diviseurs de zéro (ce sont les classes des éléments m, p de $\mathbb{Z}$ tels que $mp = n$) ; de tels anneaux ne sont donc pas des corps ; en revanche, si p est premier, $\mathbb{Z}/p\mathbb{Z}$ est un corps de caractéristique p : dans $\mathbb{Z}/3\mathbb{Z}$ par exemple, qui comporte les trois éléments $\bar{0}$, $\bar{1}$ et $\bar{2}$, $\bar{1}$ et $\bar{2}$ sont leur propre inverse pour la multiplication (voir fig. $4A_2$) ; plus généralement :

si A est un anneau commutatif et I un idéal maximal de A, alors A/I est un corps.

Citons encore les propriétés suivantes :

— tout anneau intègre unitaire fini est un corps ;

— les seuls idéaux (à gauche ou à droite) d'un corps K considéré comme anneau sont $\{0\}$ et K ; en effet, considérons un idéal à gauche I de K (le raisonnement est analogue pour un idéal à droite) ; si $I \neq \{0\}$, il existe $x \neq 0$ appartenant à I, d'où $x^{-1}x = 1 \in I$ et donc $y \cdot 1 = y$ appartient à I pour tout élément y de K : I = K.

Sous-corps

Une partie L de K qui est elle-même un corps pour l'addition et la multiplication induites par celles de K s'appelle un sous-corps de K ; on dit alors que K est un *surcorps* ou une *extension* du corps L. Un sous-corps L de K est une partie de K qui vérifie les propriétés suivantes :

— L est un sous-anneau de K (K étant considéré comme anneau) ;

— les inverses dans K de tous les éléments de L appartiennent à L.

$\mathbb{Q}$ est un sous-corps de $\mathbb{R}$, $\mathbb{C}$ une extension de $\mathbb{R}$. Tout sous-corps d'un corps K distinct de K est appelé *sous-corps propre* de K. Un corps est dit *premier* s'il ne contient aucun sous-corps propre : si p est premier, $\mathbb{Z}/p\mathbb{Z}$ est un corps premier. L'intersection d'une famille quelconque de sous-corps de K est encore un sous-corps de K ; en particulier, l'intersection de tous les sous-corps contenant une partie non vide B de K est un sous-corps de K : c'est le sous-corps de K *engendré* par B.

Homomorphismes de corps

Les corps étant des anneaux particuliers, un homomorphisme d'un corps K dans un corps K′ est simplement un homomorphisme d'anneau de K dans K′ ; mais ces homomorphismes possèdent une propriété qui nous donne des résultats particulièrement simples.

En effet, nous savons que le noyau d'un homomorphisme d'un anneau A dans un anneau A′ est un idéal bilatère de A ; comme les seuls idéaux d'un corps sont $\{0\}$ et le corps lui-même, si f est un homomorphisme d'un corps K dans un corps K′, deux cas peuvent se présenter :

— $f^{-1}(0) = K$ et $f(K)$ est un anneau réduit à 0 (pour tout élément x de K, $f(x) = 0$) ;

ou bien

— $f^{-1}(0) = \{0\}$; f est alors un homomorphisme injectif de K dans K′ et $f(K)$ est un corps isomorphe à K ; en particulier, si f est surjectif, les corps K et K′ sont isomorphes.

Corps des fractions d'un anneau intègre

Nous avons vu qu'un corps est dépourvu de diviseurs de zéro ; de même, tout sous-anneau de ce corps possède cette propriété d'où il résulte que tcut sous-anneau d'un corps commutatif est un anneau intègre. Nous allons voir que réciproquement, étant donné un anneau intègre A, il est possible de trouver un corps commutatif K dont A soit un sous-anneau ; on dit encore que l'on va *plonger* A dans un corps K.

Soit donc A un anneau intègre, notons A^* l'ensemble des éléments non nuls de A ; la relation :

$$xy' = x'y$$

définit une relation d'équivalence $\mathcal{R}$ dans l'ensemble $A \times A^*$:

$$(x, x') \, \mathcal{R} \, (y, y') \text{ si, et seulement si } xy' = x'y ;$$

on peut définir de plus dans $A \times A^*$, une addition par :

$$(x, x') + (y, y') = (xy' + x'y, x'y')$$

et une multiplication par :

$$(x, x') \cdot (y, y') = (xy, x'y')$$

On vérifie facilement que ces lois sont associatives et commutatives, que la multiplication est distributive par rapport à l'addition, et que, de plus, $\mathcal{R}$ est compatible avec ces lois. On peut donc définir dans l'ensemble quotient $K = A \times A^*/\mathcal{R}$ l'addition et la multiplication habituelles qui sont, si l'on convient de noter x/x' la classe d'équivalence $\overline{(x, x')}$:

$$x/x' + y/y' = (xy' + x'y)/x'y'$$
$$(x/x') \cdot (y/y') = xy/x'y'$$

— en outre, la multiplication a un élément neutre, y'/y' noté 1 (y' est un élément quelconque de A^* : pour tout $z' \in A^*$, $z'/z' = y'y'$) car :

$$(x/x') \cdot (y'/y') = (y'/y') \cdot (x/x') = x/x' \; (y'xx' = y'x'x)$$

— l'addition a un élément neutre, $0/y'$ noté $\bar{0}$ (pour tout y' de A^*, tout z' de A^*, on a $0/y' = 0/z'$) car :

$$x/x' + 0/y' = xy'/x'y' = (x/x') \cdot \bar{1} = x/x'$$

— tout élément x/x' a un opposé pour l'addition, $(-x)/x' : x/x' + (-x)/x' = 0/xx' = \bar{0}$

— tout élément x/x' distinct de $\bar{0}$ a un inverse pour la multiplication, x'/x (si $x/x' \neq \bar{0}$, alors $x \in A^*$) :

$$(x/x') \cdot (x'/x) = xx'/xx' = \bar{1}$$

K est donc bien un corps commutatif ; ce corps ne contient pas A mais, si l'on considère la partie B de K comportant les éléments de la forme xx'/x' où $x \in A$, on voit qu'elle constitue un sous-anneau de K isomorphe à A (xx'/x correspondant à x par cette bijection) ; en identifiant B à A, on obtient bien un corps commutatif K′, isomorphe à K qui contient A. K′ s'appelle le *corps des fractions* de A.

Si l'on prend pour A l'anneau $\mathbb{Z}$, le corps K′ n'est autre, par construction, que le corps $\mathbb{Q}$ des nombres rationnels.

BIBLIOGRAPHIE

BOURBAKI N., *Éléments de mathématiques,* première partie, livre II : *Algèbre,* Paris, Hermann, 1951 (2e édition). - EHRESMANN C., *Catégories et Structures,* Paris, Dunod, 1964. - GODEMENT R., *Cours d'algèbre,* Paris, Hermann, 1966. - LANG S., *Algebra,* Addison-Wesley Publishing Company (1re édition, 1969). - QUEYSANNE M., *Algèbre,* Paris, Armand Colin, 1964. - QUEYSANNE M. et DELACHET A., *l'Algèbre moderne,* « Que sais-je ? », no 661, Paris, P.U.F., 1960. - VAN DER WAERDEN, *Moderne Algebra,* Berlin Springer, 1943 (3e édition en allemand) ; traduction anglaise, New York, Vujar, 1949.

▶ *Le* chevalier de Méré *(v. 1607-1685) ; il connut* Pascal *vers 1652.*



COMBINATOIRE

Tout enfant sorti des classes primaires est censé savoir compter. Compter les billes contenues dans un sac, déterminer le nombre de voix obtenues par un candidat lors d'une élection sont des opérations simples — fastidieuses peut-être ; mais elles ne posent guère de problèmes de *méthode*. Il s'agit uniquement de procéder à une *énumération* en prenant toutes les billes du sac ou tous les bulletins de l'urne un à un.

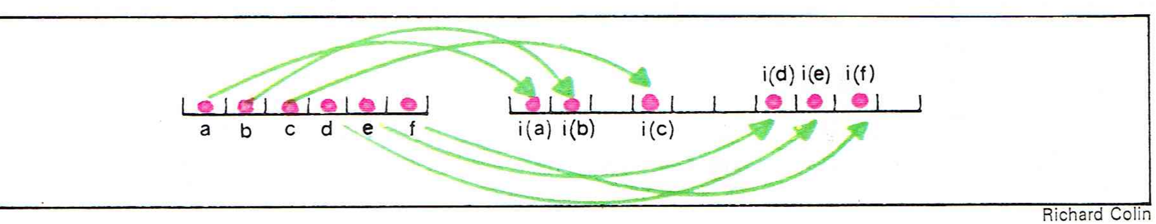
Par contre, évaluer le nombre d'anagrammes que l'on peut réaliser avec les lettres du mot *« Anticonstitutionnellement »,* calculer combien il y a de donnes possibles dans un jeu de tarots contenant les 3 bouts sont des opérations qui deviennent inextricables si l'on ne dispose pas de méthodes de dénombrement qui permettent de faire l'économie d'une énumération directe de toutes les possibilités. La combinatoire est justement l'étude de ces méthodes de dénombrement. Elle est à la base du calcul des probabilités : historiquement, elle est née d'un problème posé à Pascal par le chevalier de Méré (1654). L'auteur des *Provinciales* sut le résoudre et fournit ainsi les premiers éléments des probabilités.

La combinatoire : la mathématique des grands nombres

Soit A un ensemble (fini) ; on appelle cardinal de A (noté card A) le nombre d'éléments de A. Lorsque le nombre card A semble *a priori* très élevé, il sera souvent possible de montrer mathématiquement qu'il existe une bijection entre A et un sous-ensemble particulier de l'ensemble F^E des applications d'un ensemble E dans un ensemble F. C'est pour cela que la combinatoire s'est avant tout consacrée au dénombrement des ensembles d'applications.

Exemple 1 — Arrangements

Soit à résoudre le problème suivant : de combien de façons différentes peut-on regrouper n éléments en sous-ensembles ordonnés de p éléments ($p \leqslant n$) ? Ce nombre-là est égal au cardinal de l'ensemble $\mathfrak{J}$ des applications injectives d'un ensemble E de p éléments dans un ensemble F à n éléments (schéma ci-dessous).



Richard Colin

Soit $a \in$ E. Son image par une injection peut prendre n valeurs. Une fois une valeur choisie, un deuxième élément a a, quant à lui, le choix entre $n - 1$ valeurs (la valeur prise par a lui est interdite puisque l'application est injective). Pour les deux premiers éléments, il y a $n\ (n - 1)$ possibilités, et ainsi de suite. Au bout du compte, nous obtenons

$$A_n^p = (n - 1)\ (n - 2)\ ...\ (n - p + 1).$$

Ce nombre s'appelle nombre d'arrangements de n éléments p à p.

Lorsque $p = n$, l'on obtient le nombre de *bijections* d'un ensemble E sur lui-même. Ce nombre, que l'on nomme « factorielle n », se note $n\,!$.

$$n\,! = 1 \times 2 \times 3 \times 4 \times ... \times n$$

On convient en outre d'écrire : $1\,! = 1$; $0\,! = 1$.

Exemple 2 — Combinaisons

Si l'on convient d'identifier les arrangements qui diffèrent seulement par l'*ordre* des p éléments, on obtient l'ensemble des combinaisons de n éléments p à p. Cela revient à définir sur l'ensemble $\mathfrak{J}$ une relation d'équivalence $\mathcal{R}$: on dira que $i\,\mathcal{R}\,i'$ si, et seulement si $i\,(E) = i'\,(E)$. Le nombre de combinaisons est égal alors au cardinal de l'ensemble quotient $\mathfrak{J}/\mathcal{R}$. Toutes les classes d'équivalence de la relation $\mathcal{R}$ contiennent $p\,!$ injections : toute partie $\mathcal{E}$ de F de n éléments définit une classe d'équivalence $A\mathcal{E} = \{i \mid i\,(E) = \mathcal{E}\}$. $A\mathcal{E}$ est donc l'ensemble des bijections de $E \to \mathcal{E}$, donc card $A\mathcal{E} = p\,!$, donc

card $\mathfrak{J}/\mathcal{R} = \dfrac{\text{card}\ \mathfrak{J}}{p\,!}$; le nombre se note C_n^p (ou $\binom{p}{n}$) :

$$C_n^p = \frac{n\ (n - 1)\ ...\ (n - p + 1)}{p\,!} = \frac{n\,!}{p\,!\ (n - p)\,!}$$

on démontre les deux égalités : **(1)** $C_n^p = C_n^{n-p}$

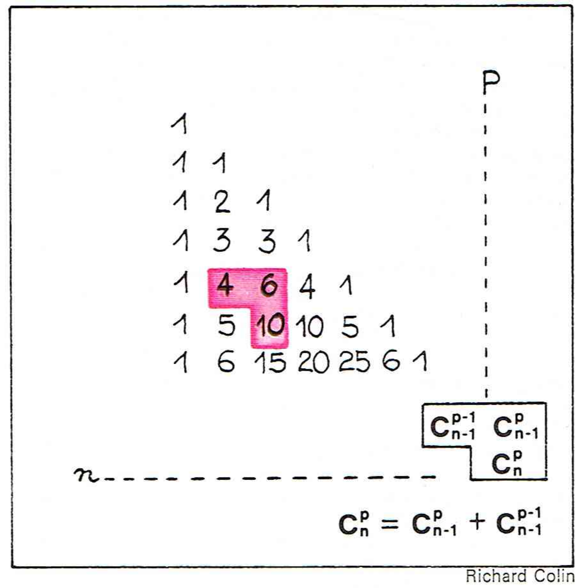
(2) $C_n^p = C_{n-1}^p + C_{n-1}^{p-1}$

Enfin le binôme de Newton $(1 + x)^n$ se développe de la manière suivante :

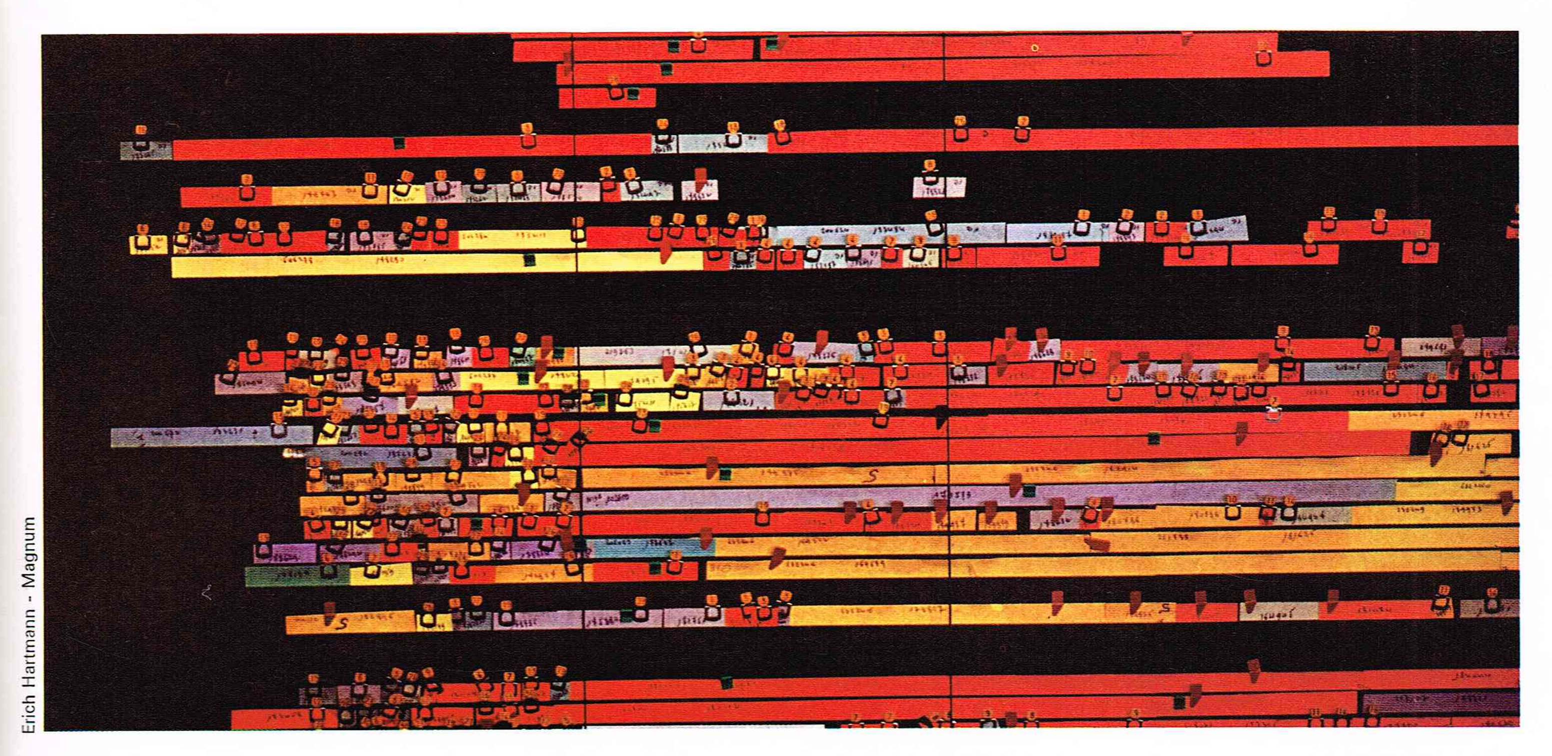
$$(1 + x)^n = 1 + nx + C_n^2\,x^2 + C_n^3\,x^3 + ... + C_n^k\,x^k + ...\,x^n$$

Les coefficients du développement peuvent être réunis dans un tableau (triangle de Pascal) que l'on construit à l'aide de l'égalité **(2)** :

▶ *Le triangle de Pascal.*



Richard Colin



Erich Hartmann - Magnum

▲ *La théorie des graphes a donné naissance à un grand nombre d'applications en echerche opérationnelle comme, notamment ici, la planification de l'écoulement de la production d'une firme automobile.*

LES GRAPHES

Un petit dessin vaut mieux qu'un grand discours. NAPOLÉON

Dans la vie courante, nous sommes souvent amenés à représenter des situations ou des problèmes par des schémas comportant des points reliés par des flèches : ainsi en est-il d'une carte routière, d'un réseau électrique ou de l'organigramme des services d'une firme. A l'évidence, ce mode de représentation possède, dans le domaine industriel, militaire ou domestique, des vertus pratiques et pédagogiques incontestables; plus encore, on s'est rendu compte qu'il pouvait faciliter la résolution de certains problèmes : en 1736, Euler étudia comment il était possible de parcourir les sept ponts de la ville de Kœnigsberg sans franchir deux fois le même. En montrant l'impossibilité du problème, il dégagea les premiers éléments de ce qui constitue aujourd'hui la *théorie des graphes*. Pour parler néanmoins de la théorie des graphes, il faudra attendre le XX[e] siècle, le développement de la théorie des ensembles et la systématisation de la démarche axiomatique : deux ouvrages fondamentaux doivent à cet égard être mentionnés : celui de D. Kœnig (Leipzig, 1936) et celui de C. Bergé (Paris, 1957).

La théorie des graphes constitue une des branches des mathématiques les plus florissantes : elle suscite l'intérêt de nombreux mathématiciens : il existe en effet beaucoup de conjectures non démontrées (dont la résolution semble d'autant plus délicate que l'énoncé est simple), et chaque année voit arriver une moisson abondante et internationale de résultats qui tentent de les résoudre.

Par ailleurs, la théorie des graphes a largement débordé sur d'autres branches des mathématiques. On en étudie aussi les applications aux groupes finis, aux catégories, ou en topologie. Les graphes ont aussi donné naissance à un grand nombre d'applications en recherche opérationnelle : les problèmes de transport, d'ordonnancement des tâches, de planning, de stocks, sont traités désormais à l'aide de méthodes et d'algorithmes (non combinatoires) qui en sont directement issus (algorithme du flot maximal de Ford et Fulkerson, méthode P. E. R. T., etc.). Là encore, les progrès réalisés au cours des vingt dernières années ont été importants, suscités par l'émergence des problèmes d'organisation et de planification qui se posent actuellement aux vastes structures administratives et industrielles, rendus possibles par le développement des ordinateurs et de l'informatique.

Définitions et propriétés générales

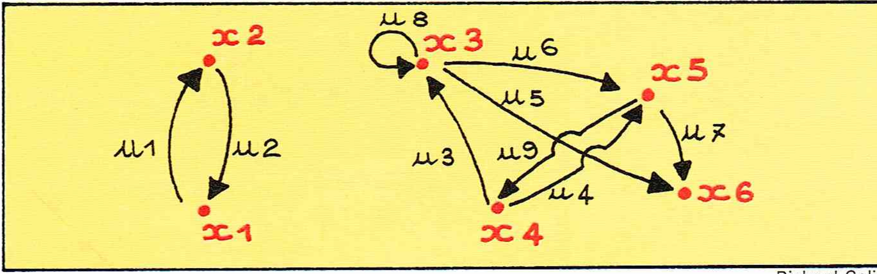
On appelle *graphe* le couple G = (X, **U**) formé par un ensemble fini X d'éléments appelés « *sommets* » et par une famille **U** de couples de sommets que l'on convient d'appeler « *arcs* ».

$$X = \{x_1, x_2, ..., x_n\}$$
$$\mathbf{U} = \{u_1, u_2, ..., u_m\}$$
$$\mathbf{U} \in X \times X$$

Un graphe est en général orienté : c'est-à-dire que l'on distingue l'arc (x_i, x_j) de l'arc (x_j, x_i) ; on suppose ici qu'entre deux sommets x_i, x_j il existe *au plus* un arc (x_i, x_j). Si l'on ne tient pas compte de l'orientation des arcs, on parlera de graphe *simple* (et d'arêtes plutôt que d'arcs). On peut ainsi définir un graphe par G = (X, Γ) où Γ est une *multiapplication* X ⇒ X (c'est-à-dire une *application* de $X \to \mathfrak{P}(X)$). Il est clair que **U** est alors le « graphe », au sens usuel de l'analyse, de Γ :

$$\mathbf{U} = \{(x, y) \mid x \in X, y \in \Gamma(x)\}$$

Exemple :



Richard Colin

le graphe comporte 6 sommets et 9 arcs (il n'est pas connexe).

$$X = \{x_1, x_2, x_3, x_4, x_5, x_6\}$$
$$\mathbf{U} = \{(x_1, x_2), (x_2, x_1), (x_4, x_3), (x_4, x_5), (x_3, x_6), (x_3, x_5), (x_5, x_6), (x_3, x_3), (x_5, x_4)\}$$

(ou $\mathbf{U} = \{u_1, u_2, u_3, ..., u_9\}$).

Un arc tel que (x_3, x_3) s'appelle une *boucle*.

Matrice boolienne d'un graphe

Il s'agit d'une matrice a_{ij} carrée d'ordre *n* définie par

$a_{ij} = 1$ si $(x_i, x_j) \in \mathbf{U}$
$= 0$, sinon.

La matrice du graphe donné en exemple est :

	1	2	3	4	5	6
1	0	1	0	0	0	0
2	1	0	0	0	0	0
3	0	0	1	0	1	1
4	0	0	1	0	1	0
5	0	0	0	1	0	1
6	0	0	0	0	0	0

L'indice de ligne correspond au sommet d'origine, l'indice de colonne au sommet extrémité.

Quelques définitions

Considérons un arc (x_i, x_j). Si x_i est appelé extrémité initiale, x_j extrémité finale, x_i est le prédécesseur de x_j, x_j le successeur de x_i. Deux *arcs* seront dits *adjacents* s'ils ont une extrémité commune (par exemple (x_4, x_5) et (x_3, x_5)). Deux *sommets* sont dits *adjacents* s'ils sont distincts et reliés par un arc.

Un *chemin* est une suite *ordonnée* d'arcs tels que l'extrémité finale de chaque arc coïncide avec l'extrémité initiale du suivant; exemple : (x_4, x_3) ; (x_3, x_5) ; (x_5, x_6).

Un *circuit* est un chemin dont le sommet initial coïncide avec le sommet terminal; exemple :

$$(x_4, x_3) ; (x_3, x_5) ; (x_5, x_4).$$

Une *chaîne* est une suite ordonnée d'arcs adjacents; exemple : (x_3, x_6) ; (x_3, x_3) ; (x_4, x_3) (ou d'arêtes dans un graphe *simple*).

Un *cycle* est une chaîne fermée; exemple :

$$(x_3, x_6) ; (x_3, x_5) ; (x_5, x_6).$$

On imposera de plus qu'il n'utilise pas deux fois le même arc.

Soit A une partie de X. On notera $\omega^+ (A)$ (respectivement $\omega^- (A)$) l'ensemble des arcs ayant l'extrémité initiale (seulement) dans A (respectivement : extrémité finale).

$$\omega^+ (A) = \{(x_i, x_j) \mid x_i \in A, x_j \notin A\}$$
$$\omega^- (A) = \{(x_i, x_j) \mid x_i \notin A, x_j \in A\}$$
$$\omega (A) = \omega^+ (A) \cup \omega^- (A)$$

par exemple : si $A = \{x_4, x_3\}$

$$\omega^+ (A) = \{u_4, u_5, u_6\} \quad \text{et} \quad \omega^- (A) = \{u_9\}$$

Recherche d'un flot maximal dans un réseau de transport

Un réseau de transport est un graphe particulier qui permet de représenter certains problèmes pratiques (problèmes de transports, d'affectation des tâches). Un réseau de transport est un graphe G = (X, **U**) tel que :

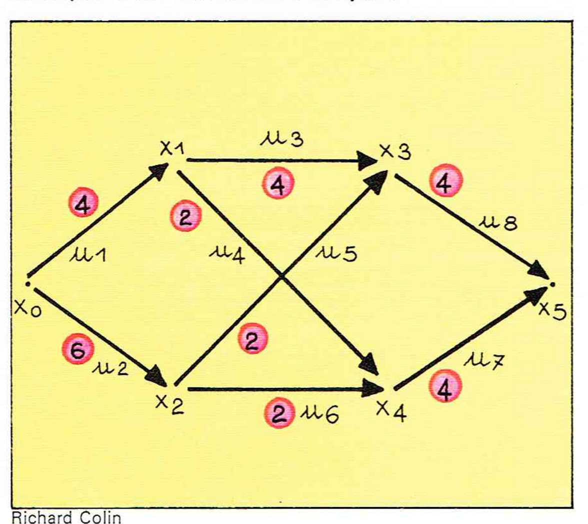
— il existe un sommet unique x_0 (appelé l'*entrée* du réseau) qui n'a pas de prédécesseur;

— il existe un sommet unique x_n *(la sortie)* qui n'a pas de successeur;

— à chaque arc est associé un nombre appelé sa *capacité*. On notera C_{ij} ou C_k la capacité de l'arc $u_k = (x_i, x_j)$;

— on supposera, pour simplifier, qu'entre deux sommets x_i, x_j du graphe, il existe au plus un arc.

Exemple d'un réseau de transport



Richard Colin

▶ ***Exemple schématique d'un* réseau de transport.**

Un réseau est donc à l'image d'un réseau électrique reliant une entrée (pôle + d'un accumulateur) à une sortie (pôle —). Chaque arc comporte des résistances, des capacités, des selfs, et l'on peut chercher à déterminer le *flux de courant* traversant le réseau, sachant qu'il n'y a pas d'accumulation d'électricité en un sommet autre que x_0 et x_n : la quantité d'électricité qui arrive en un sommet doit être égale à celle qui en part (loi de Kirchhoff); d'autre part, la quantité s'écoulant dans un arc est au plus égale à la capacité dudit arc.

On définit donc ainsi un flot dans un réseau : on appelle flot $\vec{\Phi}$ *dans un réseau* (X, **U**) possédant *m* arcs, un ensemble de nombres $\vec{\varphi} = (\varphi_1, \varphi_2, ..., \varphi_m)$, le nombre φ_k étant associé à l'arc *k*, tels que :

$$\text{(i)} \quad \sum_{k \in \omega^- \{x_j\}} \varphi_k = \sum_{k \in \omega^+ \{x_j\}} \varphi_k \quad \text{pour } x_j \in X - \{x_0, x_n\}$$
$$0 \leqslant \varphi_k \leqslant C_k \qquad \forall k = 1, 2, ..., m$$

La quantité de flot entrant dans les réseaux est égale à

$$\sum_{k \in \omega^+ (x_0)} \varphi_k$$

elle est évidemment égale à la quantité qui en sort :

$$\Phi = \sum_{k \in \omega^+ (x_o)} \varphi_k = \sum_{k \in \omega^- (x_m)} \varphi_k$$

la valeur commune à ces deux expressions est appelée *valeur du flot dans le réseau*.

Problème du flot maximal

Trouver dans le réseau G = (X, **U**) le flot de valeur Φ^* maximale.

Algorithme de Ford et Fulkerson de recherches d'un flot maximal dans un réseau

La méthode est une méthode itérative : on suppose que l'on connaît un flot, et on cherche à l'améliorer. Lors de la *j*-ième itération, on détermine un flot $\vec{\varphi}^{(j)} = \vec{\varphi}^{(j-1)} + \vec{\eta}^{(j)}$; la procédure s'arrête lorsque l'on ne peut plus faire augmenter la valeur du flot. La méthode est une méthode matricielle, dont voici le principe appliqué à l'exemple donné ci-dessus. Nous représentons ce réseau par sa matrice des capacités : il s'agit d'une matrice carrée (n, n) $\{C_{ij}\}$ où C_{ij} représente la capacité de l'arc (x_i, x_j) ; si l'arc (x_i, x_j) n'existe pas, on pose par convention $C_{ij} = 0$.

On part avec un flux initial $\vec{\varphi}^{(0)} = \vec{0}$. Comment l'améliorer? On voit que l'on peut passer de $x_0 \to x_1 \to x_3 \to x_5$.

1re itération

Au maximum, on fait passer 4 unités de $x_0 \to x_5$ par ce chemin.

On pose $\Phi^{(1)} = 4$ et $\vec{\varphi}^{(1)} = (4, 0, 4, 0, 0, 0, 0, 4)$.

On va donc modifier la matrice de manière à tenir compte du passage de flot $\vec{\varphi}^{(1)}$ [voir $M^{(0)}$]. Les capacités des arcs (x_0, x_1), (x_1, x_3) et (x_3, x_5) deviennent nulles. Par contre, celles des arcs (x_1, x_0), (x_3, x_1) et (x_5, x_3) deviennent égales à 4. (En effet, on peut admettre que l'on peut faire remonter une unité de x_1 à x_0 en en retenant une de celles qui passent de x_0 à x_1.)

2e itération

On poursuit la même procédure : il est possible d'augmenter la valeur du flot $\vec{\varphi}^{(1)}$ en empruntant le chemin (x_0, x_2, x_4, x_5). Sur la matrice $M^{(1)}$, on voit que l'on peut faire passer au plus 2 unités. On obtient alors le flux :

$$\vec{\varphi}^{(2)} = (4, 2, 4, 0, 0, 2, 2, 4)$$

et $\Phi^{(2)} = 4 + 2 = 6$. Les capacités des arcs (x_0, x_2) (x_2, x_4) et (x_4, x_5) deviennent alors $(4, 0, 0)$, celles de (x_2, x_0), (x_4, x_2) et (x_5, x_4) : $(2, 2, 2)$.

3e itération

On augmentera le flot en suivant le « chemin »

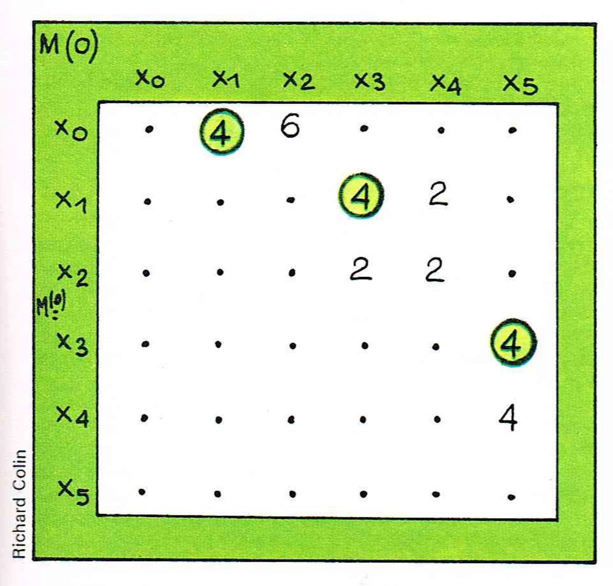
$$(x_0, x_2, x_3, x_1, x_4, x_5)$$

l'arc (x_3, x_1) n'existe pas, mais il est possible d'augmenter néanmoins le flot en détournant une unité qui va de x_1 à x_3 vers x_4. Compte tenu des capacités, on voit que le flux est amélioré de la sorte de 2 unités

$$\vec{\varphi}^{(3)} = (4, 4, 2, 2, 2, 2, 4, 4)$$
$$\Phi^3 = 8$$

Si l'on cherche à poursuivre la procédure, on ne parviendra pas à trouver un chemin composé d'arcs de capacités non nulles grâce auquel l'on puisse atteindre x_5. Donc $\vec{\varphi}^* = \vec{\varphi}^{(3)}$; $\Phi^* = 8$. On remarquera que le flot $\vec{\varphi}^*$ s'obtient en faisant la différence des matrices $M^{(0)} - M^{(3)}$.

Ford et Fulkerson ont montré que le flot $\vec{\varphi}^*$ obtenu par cette méthode est un flot maximal.



M(0)

	x_0	x_1	x_2	x_3	x_4	x_5
x_0	•	(4)	6	•	•	•
x_1	•	•	•	(4)	2	•
x_2	•	•	•	2	2	•
x_3	•	•	•	•	•	(4)
x_4	•	•	•	•	•	4
x_5	•	•	•	•	•	•

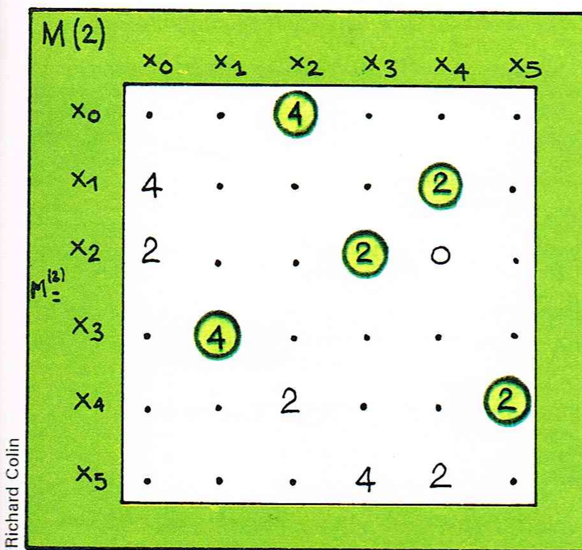
Richard Colin

M(1)

	x_0	x_1	x_2	x_3	x_4	x_5
x_0	•	o	(6)	•	•	•
x_1	4	•	•	o	2	•
x_2	•	•	•	2	(2)	•
x_3	•	4	•	•	•	o
x_4	•	•	•	•	•	(4)
x_5	•	•	•	4	•	•

Richard Colin

◀ *La matrice* $M^{(j)}$ *indique les capacités réelles du réseau, sachant que le flot* $\varphi^{(j)}$ *y circule. Les O (ou les ●) indiquent des arcs qui, soit ne sont pas parcourus par le flot, soit sont saturés par celui-ci. Lorsque, au cours d'une itération, l'on sature un arc [par exemple l'arc* (x_1, x_4)*, ou le troisième] cela revient à admettre que l'on a augmenté la capacité de l'arc contraire* (x_4, x_1) *; ce que l'on indique sur la matrice de l'itération.*



M(2)

	x_0	x_1	x_2	x_3	x_4	x_5
x_0	•	•	(4)	•	•	•
x_1	4	•	•	•	(2)	•
x_2	2	•	•	(2)	o	•
x_3	•	(4)	•	•	•	•
x_4	•	•	2	•	•	(2)
x_5	•	•	•	4	2	•

Richard Colin

M(3)

	x_0	x_1	x_2	x_3	x_4	x_5
x_0	•	•	2	•	•	•
x_1	4	•	•	2	o	•
x_2	4	•	•	o	•	•
x_3	•	2	2	•	•	•
x_4	•	2	2	•	•	o
x_5	•	•	•	4	4	o

Richard Colin

Théorème de dualité - Maxflow - Min cut

Ce théorème (1957) apporte une intéressante caractérisation d'un flot optimal : soit A ⊂ X tel que $x_0 \notin$ et $x_n \in A$, l'ensemble $\omega^-(A)$ est appelé *une coupe de capacité* $C_A = \sum_{k \in \omega^-(A)} C_k$. Dans l'exemple précédent, avec $A = \{x_3, x_5, x_n\}$, on a $C_A = 8$. Le théorème s'énonce sous la forme suivante : *Dans un réseau de transport, la valeur maximale d'un flot est égale à la capacité minimale d'une coupe.* Dans l'exemple, $\Phi^* = C_A$ et toute coupe du graphe a une capacité supérieure à celle de A.

Chemins critiques — Méthode P.E.R.T.

Considérons un entrepreneur chargé de construire une maison. Il s'agit d'un projet dont la réalisation nécessite l'accomplissement de plusieurs opérations (ou tâches) :

A = creusage des fondations,
B = construction du gros œuvre,
C = mise en place de l'installation électrique,
D = du chauffage,
E = réalisation des peintures extérieures,
F = des peintures intérieures.

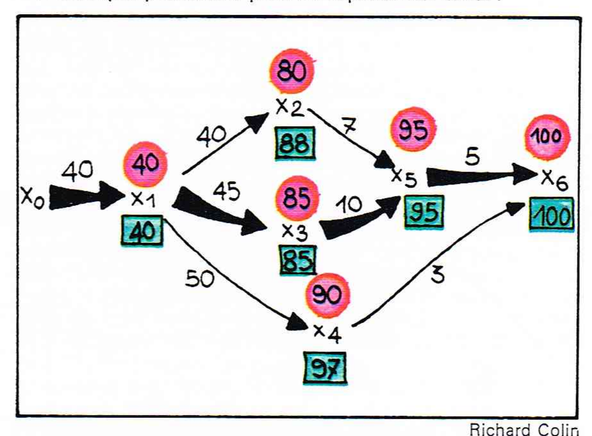
Toutes ces opérations sont soumises à certaines contraintes temporelles : par exemple, les fondations doivent être terminées avant de commencer le gros œuvre ; il est nécessaire d'avoir fini de monter le circuit électrique avant de commencer les peintures intérieures. Par contre, il est possible de réaliser simultanément les peintures extérieures et l'installation du chauffage. On représente l'ensemble des travaux par un graphe construit ainsi :

— à chaque opération est associé un *arc ;*

— les sommets du graphe correspondent à des *étapes* marquant le début de certaines opérations et la fin d'autres ;

— il existe un sommet « début des travaux » et un sommet « fin des travaux ».

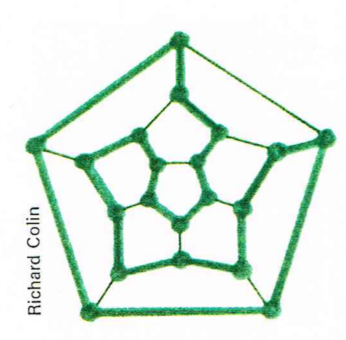
L'exemple précédent peut se représenter ainsi :



Richard Colin

◀ *Graphe représentant l'ensemble des travaux entrepris lors de la construction d'une maison.*

x_0 = début des travaux
x_1 = début du gros œuvre
x_2 = début de l'installation de l'électricité
x_3 = début de l'installation du chauffage
x_4 = début des peintures extérieures
x_5 = début des peintures intérieures
x_6 = fin des travaux.



Richard Colin

▲ ***Polyèdre à 12 faces pentagonales et à 20 sommets illustrant l'exemple du* jeu hamiltonien *(voir texte).***

La durée de chaque tâche est évaluée préalablement et figure sur le graphe (en jours). Attribuons à x_0 la date 0. Nous cherchons à évaluer la date à laquelle seront achevés les travaux : pour cela nous déterminons, par chaque tâche, la *date au plus tôt à laquelle* elle peut être réalisée, compte tenu de ses relations d'antériorité.

Soit t_{ij} la durée de la tâche (x_i, x_j). On calcule, pour chaque sommet x_i, la date t_i, date au plus tôt à laquelle on doit être en x_i. On calcule les t_i de proche en proche en partant de x_0 ($t_0 = 0$) par la formule

$$t_i = \max_h (t_h + t_{ij}) \; \forall h \mid (x_h, x_i) \in \omega^-(x_i).$$

Dans l'exemple $t_0 = 0, t_1 = 40, t_2 = 80, t_3 = 85, t_4 = 90,$

$$t_5 = \max\{t_3 + 10, t_2 + 7\} = 95,$$
$$t_6 = \max\{t_5 + 5, t_4 + 3\} = 100.$$

En effet, on ne peut réaliser la tâche (x_5, x_6) *avant* que la plus longue des deux tâches précédentes (x_2, x_5) et (x_3, x_5) ne soit terminée. Il est clair que la durée totale de la construction est donnée par des

$$t = 100.$$

Pour que la durée totale ne dépasse pas 100, il faut ne pas commencer les peintures intérieures après

$$t_5^* = 100 - 5 = 95$$

ni les peintures extérieures après $t_4^* = 100 - 3 = 97$. On détermine ainsi la *date au plus tard* de réalisation des tâches :

$$t_2^* = 95 - 7 = 88, \quad t_3^* = 95 - 10 = 85,$$
$$t_1^* = \min\{88 - 40, 85 - 45, 97 - 50\} = 40, \; t_0^* = 0.$$

En effet, il ne faut pas terminer la tâche (x_0, x_1) après 40 jours, sinon la durée totale de la construction dépassera 100 jours : on calcule ainsi les t_i^* de proche en proche en partant de x_n (avec $t_n = 100$) par la formule :

$$t_i^* = \min_h (t_h - t_{ij}) \quad \forall h \mid (a_i, a_h) \in \omega^-(a_i).$$

Recherche du chemin critique

Le chemin critique est formé des sommets pour lesquels il y a coïncidence de leur date au plus tard et de leur date au plus tôt : $t_i = t_i^* \Leftrightarrow x_i$ appartient *au chemin critique*.

Ces étapes sont à surveiller étroitement : elles sont critiques dans la mesure où le moindre retard les affectant contribue à dépasser la durée totale du projet. Dans l'exemple, le chemin critique est :

$$(x_0, x_1, x_3, x_5, x_6)$$

∗ *intervalle de flottement* $(t_i^* - t_i)$ *d'une étape* x_i *:* il est nul par les tâches critiques ;

∗ *marge totale de la tâche* (x_i, x_j) *:* à partir de la date t_i, on peut retarder le début de la tâche (x_i, x_j) de $t_j^* - t_i - t_{ij}$ sans augmenter la durée totale des travaux.

L'intérêt de la **méthode P. E. R. T.** *(Program Evaluation Research Task)* réside en ce qu'elle permet de déterminer les tâches critiques. Il est alors possible de chercher à diminuer la durée totale du projet en tentant de décomposer les tâches critiques en opérations dont certaines pourront être réalisées simultanément avec d'autres tâches. Dans l'exemple, si on sépare la tâche « chauffage » en deux opérations : pose des radiateurs et des tuyaux d'une part, et construction de la cave, pose de la chaudière d'autre part (tâches (x_1, x'_3) et (x_1, x''_3)), on obtient un nouveau graphe, ci-dessous, (sachant que l'on peut réaliser les peintures intérieures, dès que les radiateurs et les tuyaux sont posés) :

Richard Colin

x0 40 x1 (40) [40]
40 x2 (80) [83] 7
45 x'3 (85) [86] 4
x5 (89) [90] 5 x6 [95]
45 x''3 (85) [85] 10
50 x4 (90) [92] 3

La durée totale du projet a été réduite à 95 jours. Mentionnons enfin qu'il existe d'autres méthodes pour déterminer le chemin critique d'un graphe : la méthode des potentiels, qui ne nécessite pas le tracé du graphe, ou bien encore la méthode de Bellman qui repose sur la technique et les principes de la programmation dynamique.

Problèmes hamiltoniens

Soit $G = (X, \mathbf{U})$ un graphe comprenant n sommets. On dit qu'un *chemin* $\mu = [x_1, x_2, ..., x_n]$ est *hamiltonien* s'il passe une fois et une seule par chaque sommet du graphe. De même, on dira qu'un *circuit* $\mu = [x_1, x_2, ..., x_1]$ est hamiltonien s'il passe une fois et une seule par chaque sommet du graphe. Dans un graphe *simple*, on définit de même une *chaîne* hamiltonienne ou un *cycle* hamiltonien.

Exemple : W. Hamilton proposa en 1859 le jeu suivant : soit 20 villes réparties sur le globe terrestre : $a, b, c, ..., t$ que l'on représente pour simplifier par les sommets d'un dodécaèdre régulier (polyèdre à 12 faces pentagonales et à 20 sommets). On se propose de passer une fois et une seule par chacune de ces villes et de revenir à son point de départ en utilisant les seules arêtes du dodécaèdre. Ce jeu revient à chercher un cycle hamiltonien par le graphe de la figure ci-contre dans la marge (le problème étudié par Euler dans les 7 ponts de Kœnigsberg est un problème analogue).

En recherche opérationnelle, on est souvent amené à résoudre des problèmes d'ordonnancement : un certain nombre de tâches sont à réaliser : on construit alors le graphe dont les sommets représentent les tâches : deux sommets x_i et x_j sont alors reliés par un arc s'il est possible de réaliser x_i *avant* x_j. Un ordonnancement des tâches est ainsi représenté par un chemin hamiltonien. La recherche d'un tel chemin (d'un circuit, d'une chaîne ou d'un cycle) est ce que l'on appelle un problème hamiltonien.

Nous nous restreindrons ici à l'étude des cycles hamiltoniens dans un graphe *simple*. Il s'agit à ce jour d'un problème ouvert puisque l'on n'a pas encore réussi à trouver des conditions nécessaires et suffisantes d'existence d'un cycle hamiltonien. L'on ne possède que des conditions suffisantes qui toutes procèdent de la remarque suivante : soit x un sommet d'un graphe, on appelle degré (et on note $d(x)$) le nombre d'arcs incidents (intérieurement et extérieurement) au sommet x. Il semble intuitif que, plus les degrés sont élevés, plus y il a de chances pour que le graphe admette un cycle hamiltonien. Depuis Dirac (1952), on a vu apparaître dans la littérature plusieurs théorèmes d'existence utilisant la séquence des degrés de tous les sommets du graphe.

Le théorème le plus récent et en un certain sens le plus général (c'est-à-dire qui réduit tous les précédents à des corollaires) date de 1971 et est dû au mathématicien V. Chvátal.

Théorème de Chvátal

Soit un graphe simple G *à* n *sommets,* $n \geqslant 3$*, dont la suite des degrés rangés par ordre non décroissant* $d_1 \leqslant d_2 \leqslant ... \leqslant d_n$ *vérifie la propriété : pour tout* $k \leqslant \frac{n}{2}$ *on a l'implication suivante :*

$$d_k \leqslant k \Rightarrow d_{n-k} \geqslant n - k.$$

Alors le graphe G *admet un cycle hamiltonien.*

Les deux théorèmes suivants sont antérieurs à celui de Chvátal. Ils s'en déduisent désormais très facilement.

Théorème d'Ore (1960)

Soit G *un graphe simple, à* n *sommets* ($n \geqslant 3$). *Si pour tout couple de sommets* (x, y) *non adjacents on a :*

$$d(x) + d(y) \geqslant n,$$

alors le graphe admet un cycle hamiltonien.

En effet, rangeons les sommets par ordre de degrés non décroissants. Si un point x_k vérifie $d_k = d(x_k) \leqslant k \leqslant \frac{n}{2}$ alors x_k possède au plus k voisins, et il y a donc $n - k$ sommets non adjacents à x_k dont le degré est, d'après l'hypothèse, supérieur ou égal à $n - k$, soit :

$$d_{n-k} \geqslant n - k.$$

D'où le résultat en appliquant le théorème de Chvátal.

Théorème de Dirac (1952)

Soit G *un graphe à n sommets, n* $\geqslant$ 3, *tel que pour tout sommet x* $d(x) \geqslant \frac{n}{2}$, *alors ce graphe admet un cycle hamiltonien.* Ceci découle du théorème d'Ore.

Autres problèmes — autres méthodes

Conjecture des quatre couleurs

Est-il possible de colorier toute carte de géographie avec quatre couleurs, de sorte que deux régions ayant une ligne de frontière commune soient de différentes couleurs ? Ce problème célèbre n'a toujours pas reçu de solution, mais il a suscité l'intérêt des mathématiciens. En fait, il s'agit d'un problème de graphe que l'on peut formuler ainsi : on appelle *nombre chromatique* d'un graphe G le plus petit nombre de couleurs nécessaires pour colorier les sommets de sorte que deux sommets adjacents distincts ne soient pas de même couleur. On le désigne par γ (G). Un graphe planaire est un graphe que l'on peut représenter sur un plan sans que deux arcs quelconques puissent se couper ailleurs qu'en un sommet.

Il existe de très nombreux travaux sur le sujet depuis 1875. Mentionnons les résultats les plus récents :

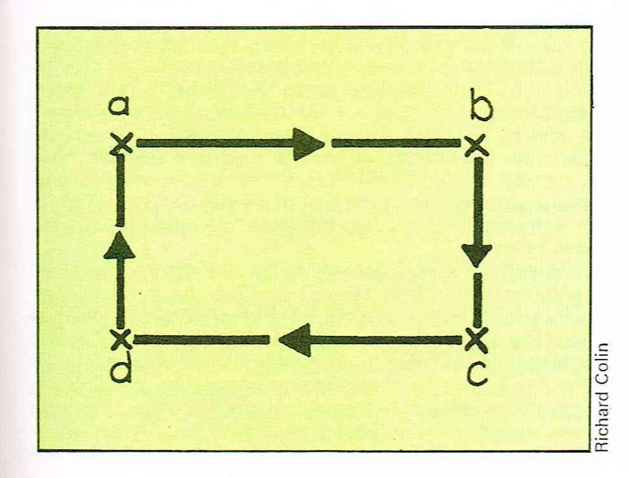
- Un graphe planaire qui ne contient pas de cycles de longueur 3 est coloriable avec trois couleurs (Grünbaum 1963).
- Si G est un graphe planaire, alors γ (G) $\leqslant$ 4.

Noyau d'un graphe

Soit un graphe G (X, Γ), on dit qu'un sous-ensemble S $\subset$ X est un *noyau* s'il vérifie les deux propriétés :

(1) $x \in S \quad \Gamma(x) \cap S = \varnothing$
(2) $x \notin S \quad \Gamma(x) \cap S \neq \varnothing$

ce graphe (ci-dessous) admet deux noyaux $\{a, c\}$ et $\{b, d\}$.



Richard Colin

Un graphe peut représenter un ordre partiel sur X,

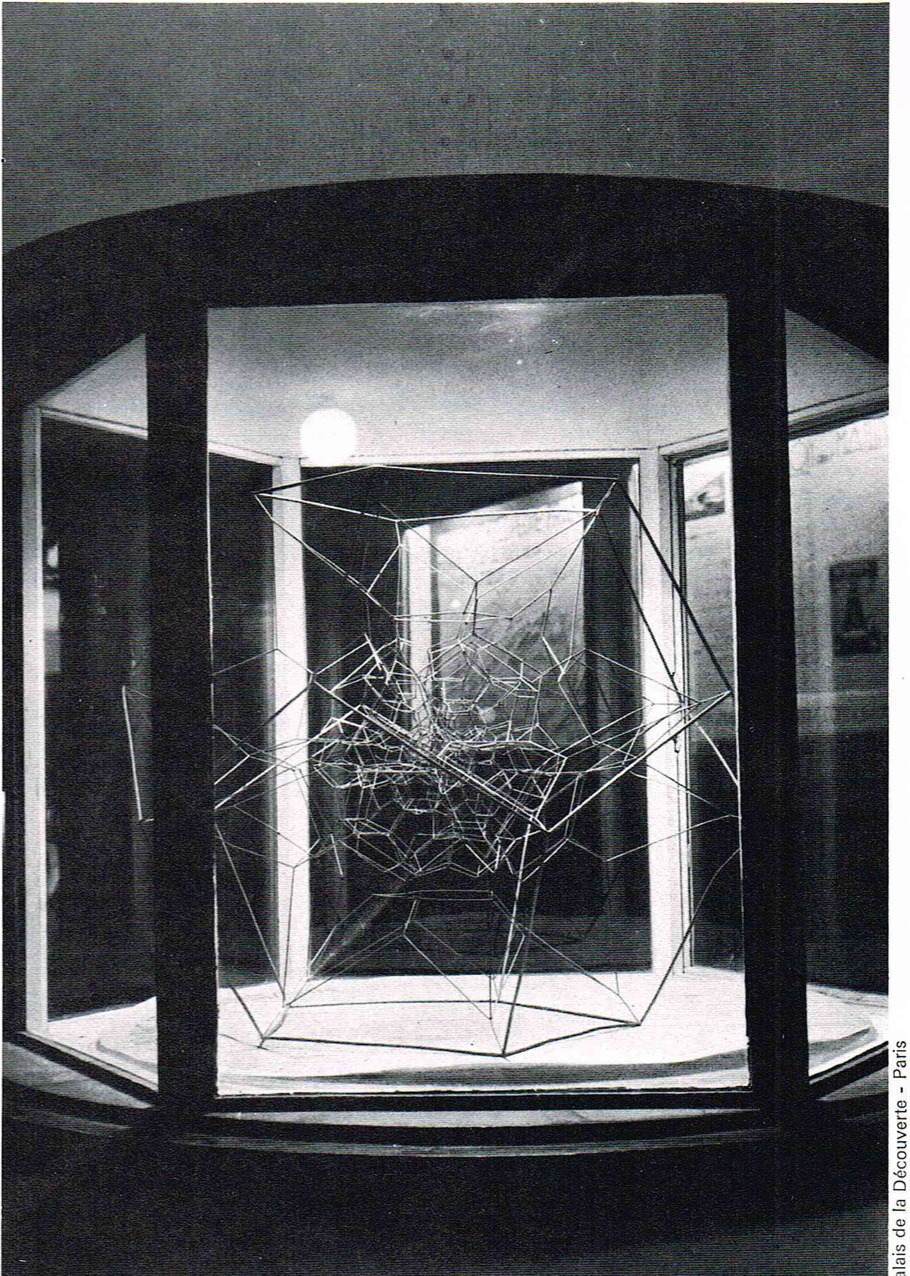
$$x \leqslant y \Leftrightarrow y \in \Gamma(x)$$

S est donc alors un ensemble d'éléments non comparables entre eux tels que tout élément pris en dehors soit préféré par au moins un élément de S.

Cette notion permet de résoudre certains jeux (jeux dits de Nim), notamment le jeu popularisé par le film d'A. Resnais *l'Année dernière à Marienbad*. En associant à ce jeu un certain graphe, on montre que le joueur qui gagne est celui qui parvient à être une fois dans le noyau et à y revenir à chaque fois (c'est pour cela que celui qui joue en second peut gagner !).

BIBLIOGRAPHIE

BERGE C., *Graphes et Hypergraphes*, Dunod, 1973. - FORD L.R. et FULKERSON D.R., *Flow in Networks*, Princeton Press, 1962 ; traduction française, Gauthier-Villars, 1967. - KAUFMAN A., *Introduction à la combinatoire en vue des applications*, Dunod, 1969. - ROY B., *Algèbre moderne et Théorie des graphes*, t. 1 et t. 2, Dunod, 1970. - SACHE A., *la Théorie des graphes*, « Que sais-je ? », P.U.F., 1974.



Palais de la Découverte - Paris

▲ ***Représentation à trois dimensions d'un solide à quatre dimensions.***

ALGÈBRE LINÉAIRE

L'algèbre linéaire est, dans le cadre actuel de son développement, une synthèse de plusieurs courants. Dans des domaines aussi divers que le calcul sur les nombres, la résolution des équations, les problèmes de géométrie classique (à résoudre tant par des méthodes synthétiques qu'analytiques), les équations différentielles, etc., tous les mathématiciens ont eu à traiter des problèmes linéaires et ont contribué à l'évolution de la méthode mathématique vers ce que l'on peut appeler la *pensée linéaire*.

Car il s'agit bien d'une forme de pensée mathématique, et il est clair maintenant que le concept de linéarisation d'un problème est très fécond. Les propriétés abondantes et des plus diversifiées que l'on peut trouver pour tout ce qui satisfait au cadre linéaire donnent à toute approximation de cette forme un très grand intérêt.

Il faut voir que cette notion est des plus anciennes et que, depuis la règle de trois jusqu'à la géométrie différentielle, en passant par la linéarisation de la théorie de Galois ou même l'étude des fonctions elliptiques, on a toujours — plus ou moins consciemment — résolu des problèmes linéaires.

On peut mentionner que les applications couvrent tous les domaines.

— Sur le plan local, car, depuis l'idée classique de la pente de la tangente à une courbe, égale à la valeur de la dérivée, jusqu'à la notion d'application linéaire dérivée dans un espace de Banach, on a approximé une fonction par une application linéaire ou affine (ainsi encore des courbes définies comme enveloppes de leurs tangentes).

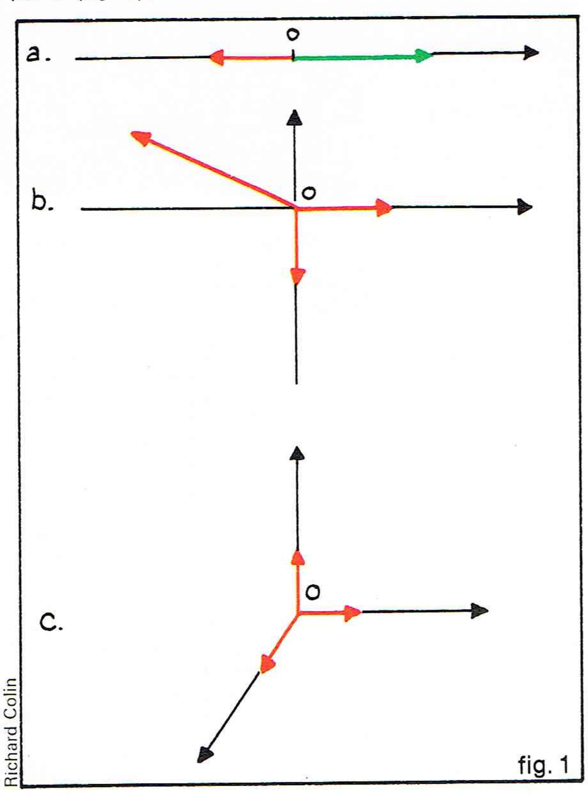
— Sur le plan global, car on cherche souvent à mettre un problème dans un cadre linéaire : c'est par exemple la notion moderne de distribution qui généralise celle de fonction, ou mieux encore celle de mesure.

De Grassmann (1809-1877) à Hilbert (1862-1943), de nombreux mathématiciens ont contribué à l'élaboration formelle de ce puissant outil. On peut ainsi citer : Möbius (1790-1868), Hamilton (1805-1865), Sylvester (1814-1897), Weierstrass (1815-1897), Cayley (1821-1895), Hermite (1822-1901), Kronecker (1823-1891), Lie (1842-1899), et bien d'autres encore.

Description d'un espace vectoriel — Bases

La structure d'un espace vectoriel est déterminée par les opérations que l'on peut réaliser sur ses éléments, c'est-à-dire le produit par un élément du corps de base (scalaire) et l'addition. L'associativité de l'addition permet de combiner encore plus globalement les éléments d'un espace vectoriel E par l'intermédiaire d'éléments du corps de base K. Soit $(x_n)_{n \in \mathbb{N}}$ une suite d'éléments de E; on appelle *combinaison linéaire* des x_n toute expression de la forme $\sum_{n \in \mathbb{N}} \alpha_n x_n$ où $\alpha_n \in$ K, et telle que seulement un nombre fini des α_n soit différent de zéro. Dans ce cas, on a en fait une somme finie de termes, tous éléments de E. Le résultat a donc bien un sens et est un élément de E Notre propos, maintenant, va être de montrer comment il est possible de décrire tous les éléments de E — et « aux moindres frais » (c'est-à-dire avec le minimum de choses données *a priori*) — par cette construction, grâce à la structure d'espace vectoriel sur E.

Soit donc un ensemble d'éléments de E appelé S. On désigne par {S} l'ensemble des combinaisons linéaires d'éléments de S (il faut bien noter que, même si S ne contient qu'un nombre fini d'éléments, {S} peut être infini même non dénombrable dès lors que Card K n'est pas fini), qui est bien évidemment un sous-espace vectoriel de E : on dit que S *engendre* {S}. On peut remarquer que {S} est le plus petit des sous-espaces vectoriels de E qui contiennent S. Dans le cas où {S} = E, on dit que S est un *système de générateurs* de E, ou que E est engendré par S (fig. 1).



▶ **Figure 1** — ***dans ces trois dessins, on montre successivement : a, un système générateur de 2 éléments pour ℝ (espace vectoriel sur lui-même); b, un système générateur de 3 éléments pour ℝ² (espace vectoriel sur ℝ); c, un système générateur de 3 éléments pour ℝ³ (espace vectoriel sur ℝ).***

D'autre part, il est clair que si l'on ajoute à S des éléments qui sont combinaisons linéaires d'éléments de S, on ne modifie pas {S}. Un système générateur de E reste donc générateur si on le complète par d'autres éléments de E. La question est alors de savoir s'il est possible d'en enlever mais tout en laissant à S son caractère générateur; elle est déjà presque résolue d'après ce que nous avons déjà dit puisqu'un élément « semble être en trop » dans S s'il est combinaison linéaire d'autres éléments de S.

On définit alors un *système libre* ou *indépendant linéairement* comme un système tel que la seule combinaison linéaire de ses éléments qui soit nulle soit celle où tous les scalaires sont nuls; soit :

$$\alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_p x_p \Rightarrow \alpha_1 = \alpha_2 = ... = \alpha_p = 0$$

Dans le cas contraire — le système est dit *lié* ou *linéairement dépendant* — on aurait une combinaison linéaire nulle :

$$\alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_p x_p = 0$$

où l'un, au moins, des α_i serait non nul, soit par exemple α_k, et l'on pourrait écrire :

$$x_k = \frac{-1}{\alpha_k} (\alpha_1 x_1 + ... + \alpha_{k-1} x_{k-1} + \alpha_{k+1} x_{k+1} + ... + \alpha_p x_p)$$

donc x_k serait combinaison linéaire des autres x_i.

On doit remarquer tout de suite qu'un système libre reste libre si on lui ôte certains de ses éléments, mais par contre peut perdre ce caractère si on lui en ajoute. C'est donc la situation opposée à celle des systèmes générateurs.

Si maintenant un système B, libre dans E, est tel que {B} = E (donc s'il est générateur aussi), on dit que B est une *base* de E.

Lorsque l'on peut trouver une famille finie S, génératrice de E ({S} = E), on dit que E est de *dimension finie*. Dans le cas contraire, on dit que E est de *dimension infinie*, ce que l'on note $\dim_K E = +\infty$.

Il est très facile de montrer que tout espace vectoriel de dimension finie admet une base; toutefois ce résultat ne peut être étendu au cas de la dimension infinie que par une démonstration assez délicate faisant intervenir le lemme de Zorn. Ce que l'on voit déjà, c'est qu'une base de E contient au moins autant d'éléments que n'importe quel système libre, et n'en contient pas plus qu'un quelconque système générateur. C'est bien là le compromis du « descriptif aux moindres frais » qui était cherché.

De cette propriété découle le fait que deux bases quelconques d'un même espace vectoriel E, de dimension finie, ont le même nombre d'éléments. C'est ce nombre que l'on appelle la *dimension* de E, notée $\dim_K E$ (cette notation, qui précise le corps de base de E, se simplifie en dim E lorsque aucune confusion n'est possible, le corps K ayant été fixé une fois pour toutes). Pour un espace vectoriel de dimension infinie, deux bases quelconques sont équipotentes, c'est-à-dire qu'il existe une bijection de l'une sur l'autre. Un exemple simple montre l'influence du corps K sur la dimension de E : dans l'espace vectoriel $\mathbb{R}^2$ sur le corps $\mathbb{R}$, les deux vecteurs $e_1 = (1, 0)$ et $e_2 = (0, 1)$ forment un système libre, car la combinaison linéaire

$$\alpha_1 e_1 + \alpha_2 e_2 = \alpha_1 \cdot (1, 0) + \alpha_2 \cdot (0, 1) = (\alpha_1, 0) + (0, \alpha_2) = (\alpha_1, \alpha_2)$$

ne peut être égale au vecteur nul (0, 0) que si $\alpha_1 = \alpha_2 = 0$; d'autre part, ils forment un système générateur car, si (x, y) est un élément quelconque de $\mathbb{R}^2$, on a bien

$$(x, y) = x e_1 + y e_2$$

La base ainsi formée est dite *base canonique* de $\mathbb{R}^2$ et on a $\dim_{\mathbb{R}} \mathbb{R}^2 = 2$. Toutefois, on sait que $\mathbb{R}^2$ possède une structure de corps, donc d'espace vectoriel sur lui-même, si l'on y définit le produit :

$$(\alpha, \beta) \cdot (x, y) = (\alpha x - \beta y, \alpha y + \beta x)$$

Dans ces conditions, tout vecteur de $\mathbb{R}^2$ est le produit d'un vecteur fixe, à composantes non nulles, de $\mathbb{R}^2$ par un élément (cette fois pris comme scalaire) de $\mathbb{R}^2$. Puisqu'un système formé d'un seul vecteur non nul est libre, on en déduit que toute base de $\mathbb{R}^2$, espace vectoriel sur lui-même, est formée par un seul élément et que $\dim_{\mathbb{R}^2} \mathbb{R}^2 = 1$.

Les notions de base et de dimension sont donc liées à la structure de l'espace vectoriel considéré E (donc entre autres au corps de base choisi), et non seulement aux éléments, non structurés, de E. La propriété fondamentale d'une base est donnée par le théorème qui suit.

Théorème. Pour tout élément x d'un espace vectoriel E sur un corps K, muni d'une base B, il existe *une et une seule* combinaison linéaire d'éléments de B, à coefficients dans K, qui soit égale à x. (On dira que tout élément de E se décompose de façon unique selon les vecteurs de base.)

En effet, B étant générateur, une telle combinaison linéaire existe. D'autre part, s'il en existait deux :

$\sum_{i \in I} \alpha_i e_i = x$ et $\sum_{i \in I} \beta_i e_i = x$ où seul un nombre fini parmi les α_i comme parmi les β_i ne sont pas nuls, on pourrait écrire :

$$0 = x - x = \sum_{i \in I} (\alpha_i - \beta_i)\, e_i = \sum_{i \in I} \gamma_i e_i$$

où seul un nombre fini de coefficients γ_i est différent de zéro. Donc la combinaison linéaire $\sum_{i \in I} \gamma_i e_i$ du système libre formé par les vecteurs e_i, est nulle ; ce qui n'a lieu que si tous les termes γ_i sont nuls. Par conséquent, $\alpha_i = \beta_i$ pour tout $i \in I$. Les deux décompositions ne peuvent donc être différentes si B est libre, donc *a fortiori* si c'est une base.

Le résultat suivant est couramment utilisé pour construire une base.

Théorème (base incomplète). Soit L un système libre de E et G une partie génératrice de E ; alors il est possible de construire une base de E en « complétant » L par une partie de G.

Il est donc bien clair que, de tout système générateur, on peut extraire au moins une base.

Ce théorème est très aisé à démontrer en dimension finie mais le lemme de Zorn est encore une fois nécessaire si l'on veut étendre ce résultat au cas de la dimension infinie. On en déduit qu'un sous-espace vectoriel V d'un espace vectoriel E (sur un corps K) de dimension finie est aussi de dimension finie, et l'on a :

$$\dim_{\mathbb{K}} V \leqslant \dim_{\mathbb{K}} E\,;$$

l'égalité ne peut avoir lieu que si V = E (et l'on pose par convention que la dimension de l'espace vectoriel $\{0\}$ est nulle).

L'intersection de deux sous-espaces vectoriels est un sous-espace vectoriel, mais leur réunion n'en est pas un en général. Plus exactement, on peut définir le sous-espace vectoriel engendré par la réunion de deux sous-espaces vectoriels U et V, c'est-à-dire l'ensemble des combinaisons linéaires d'éléments appartenant soit à U, soit à V. Ce sous-espace se nomme somme de U et V et se note U + V (il est bien formé d'éléments de la forme $u + v$ où $u \in U$ et $v \in V$). Dans le cas particulier où $U \cap V = 0$, on dit que la somme est directe et l'on note $U \oplus V$.

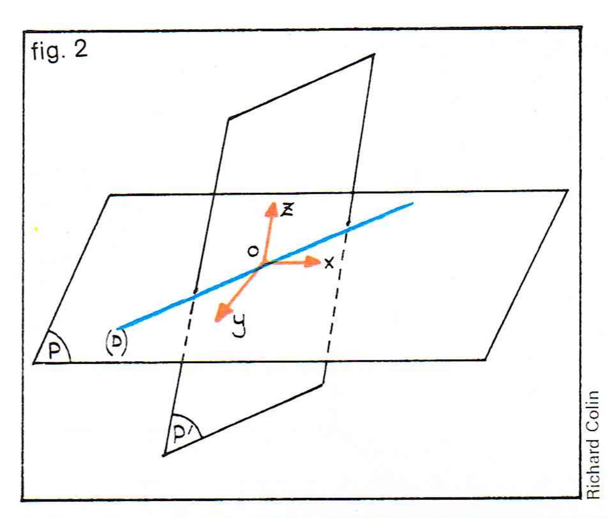
Ainsi, l'intersection dans $\mathbb{R}^3$ de deux plans passant par l'origine est une droite (s'ils ne sont pas confondus) passant par O (*fig.* 2) ; l'intersection d'un plan passant par O et d'une droite passant par O, non contenue dans ce plan, est $\{O\}$; si la droite est incluse dans le plan, l'intersection est égale à cette droite (*fig.* 3). La réunion de deux plans, par contre, n'est pas un sous-espace vectoriel. Dans $\mathbb{R}^2$, la réunion de deux droites n'est pas un sous-espace vectoriel, comme le montre la règle du parallélogramme (*fig.* 4). La somme de deux plans sécants de $\mathbb{R}^3$ n'est autre que $\mathbb{R}^3$; mais cette somme n'est pas directe. La somme d'un plan et d'une droite perpendiculaire à ce plan en O (*fig.* 5) est encore $\mathbb{R}^3$, et cette somme est directe.

Dans le cas où $E = U \oplus V$, une base de E est formée par la réunion d'une base de U et d'une base de V, et si la dimension de E est finie, on a :

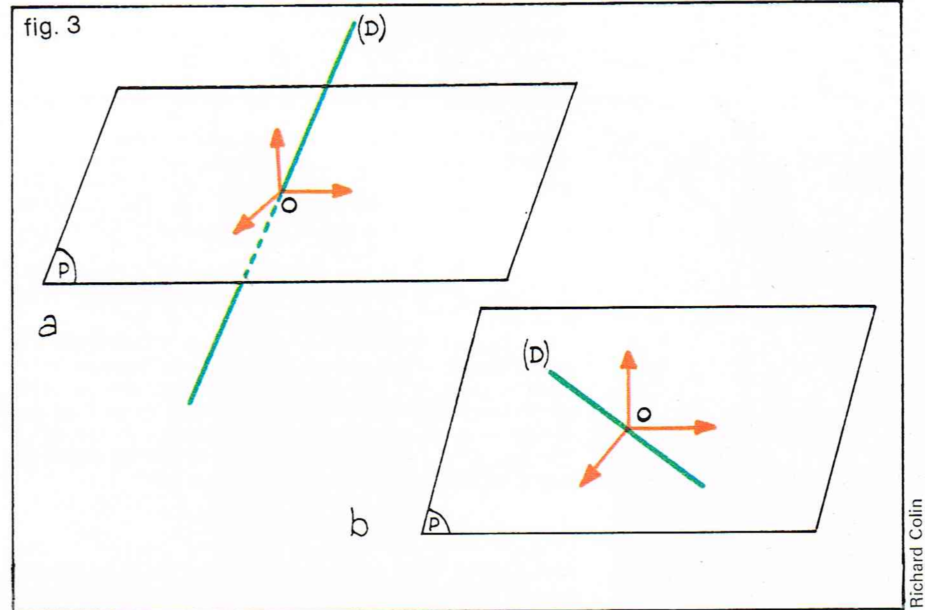
$$\dim_{\mathbb{K}} E = \dim_{\mathbb{K}} U + \dim_{\mathbb{K}} V.$$

Cette propriété se généralise par la formule :

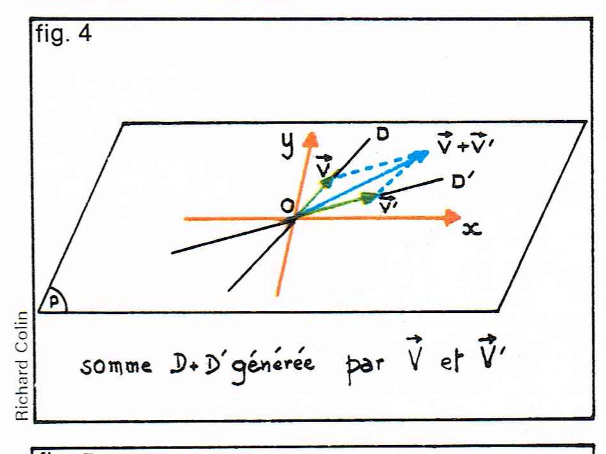
$\dim_{\mathbb{K}} (U + V) = \dim_{\mathbb{K}} U + \dim_{\mathbb{K}} V - \dim_{\mathbb{K}} (U \cap V)$ sur laquelle on ne s'étendra pas ici, mais que l'on peut vérifier sur les exemples qui suivent.

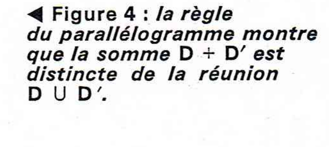


◄ **Figure 2 :** ***intersection* (D) *du plan* (P) *défini par* Ox *et* Oy *et d'un plan* (P') *passant par* Oz.**

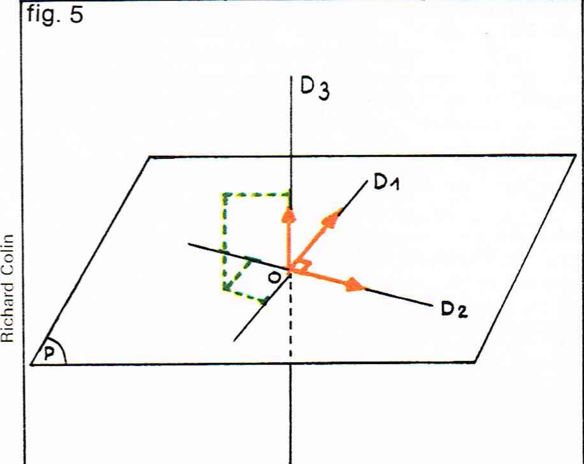


▲ **Figure 3 :** ***intersection d'un plan* (P) *et d'une droite* (D) — a, (D) *non contenue dans* (P) ; b, (D) *contenue dans* (P).**





◄ **Figure 4 :** ***la règle du parallélogramme montre que la somme* D + D' *est distincte de la réunion* D ∪ D'.**



◄ **Figure 5 :** ***l'espace* $\mathbb{R}^3$ *est obtenu comme somme directe du plan* (P) *et de la droite* D3.**

Ainsi, soit P le plan « horizontal » de $\mathbb{R}^3$, soit :

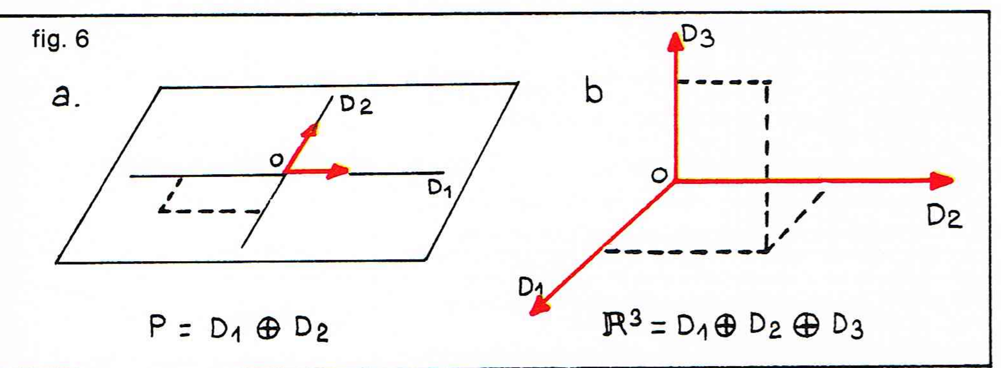
$$P = \{(x, y, z) \in \mathbb{R}^3 \text{ tels que } z = 0\}$$

et D_3 la droite « verticale », soit

$$D_3 = \{(x, y, z) \in \mathbb{R}^3 \text{ tels que } x = 0 \text{ et } y = 0\}$$

on a bien $\mathbb{R}^3 = P \oplus D_3$; en désignant par D_1, D_2, D_3 les trois axes orthonormés usuels, on a bien $P^2 = D_1 \oplus D_2$ et évidemment $\mathbb{R}^3 = D_1 \oplus D_2 \oplus D_3$ (*fig.* 6).

▼ **Figure 6 : cf. *développement dans le texte ci-contre.***



Richard Colin

On a jusqu'à présent, dans ce paragraphe, utilisé les ensembles $\mathbb{R}^2$, $\mathbb{R}^3$ et plus généralement $\mathbb{R}^n$ comme exemples. Ce n'est pas le fait d'un simple hasard : pour l'étude des espaces vectoriels de dimension finie, ces ensembles jouent un rôle fondamental que le résultat qui suit va expliciter.

Théorème. Tout espace vectoriel réel E, de dimension finie n, est isomorphe à l'espace vectoriel — sur $\mathbb{R}$ — $\mathbb{R}^n$.

Ce résultat fondamental permet de comprendre les limites des concepts utilisés selon qu'on s'intéresse à des espaces vectoriels de dimension finie ou infinie. Il se généralise : tout espace vectoriel E sur un corps K tel que $\dim_K E = n$ est isomorphe à K^n. Il provient du fait qu'un isomorphisme g d'espaces vectoriels se doit de « retranscrire » les lois additive et multiplicative, soit :

$$g(X + Y) = g(X) + g(Y)$$
$$g(a \cdot X) = a \cdot g(X)$$

donc l'image d'un vecteur quelconque de E, par g, est connue dès lors que l'on connaît les images des vecteurs d'une base donnée de E : si $v_1, v_2, ..., v_n$ désigne la base choisie dans E, tout vecteur $X \in E$ s'écrit : $X = \sum_{i=1}^{i=n} \alpha_i v_i$,

et alors

$$g(X) = \sum_{i=1}^{i=n} g(\alpha_i v_i) = \sum_{i=1}^{i=n} \alpha_i g(v_i) ;$$

la connaissance de

$$\{g(v_1), g(v_2), ..., g(v_n)\}$$

permet alors, si l'on se donne $\alpha_1, \alpha_2, ..., \alpha_n$, donc X, de déterminer g (X). Il suffit donc de pouvoir faire se correspondre, de manière biunivoque, les vecteurs d'une base de E et ceux d'une base de $\mathbb{R}^n$, ce qui est réalisable puisqu'il y en a n exactement de part et d'autre.

L'isomorphisme étant une relation d'équivalence, on peut écrire : deux espaces vectoriels de même dimension sont « identiques, à un isomorphisme près ». Ceci veut dire qu'ils sont en fait une seule et même structure d'espace vectoriel.

C'est à H. Grassmann, mathématicien allemand, que l'on doit le passage, décisif pour le développement de l'algèbre linéaire, des espaces de dimension 1, 2 et 3 — visualisables — à ceux de dimension n, de façon axiomatique claire.



Collection Viollet

▲ ***Le mathématicien anglais* Arthur Cayley *(1821-1895).***

Applications linéaires

On vient d'utiliser une application conservant les structures (donc un homomorphisme) qui, de plus, était bijective. Ce sont ces applications, conservant la structure linéaire (en ce sens que l'image d'une combinaison linéaire est la combinaison linéaire des images), qui permettent de donner un sens global au calcul matriciel tel que Cayley le formula, et à ses développements fondamentaux dus à Hermite et à Sylvester. On les appelle *applications linéaires*, et ce sont donc des applications g d'un espace vectoriel E dans un espace vectoriel F, sur un même corps K, vérifiant :

$$g(ax + by) = a\,g(x) + b\,g(y)$$

pour tous $x \in E$, $y \in E$, $a \in K$ et $b \in K$.

Quelques idées générales sur ces applications vont permettre de les utiliser, pour la résolution des systèmes d'équations linéaires entre autres.

On voit tout d'abord que $g(x - x) = g(x) - g(x)$, donc $g(0) = 0$. Mais il peut exister d'autres éléments de E ayant 0 pour image. L'ensemble de ces éléments est un sous-espace vectoriel de E, que l'on appelle *noyau* de g, ce qui se note $\mathrm{Ker}\, g$ (de la racine celtique « kernel »). On a toujours $0 \in \mathrm{Ker}\, g$; si $\{0\} = \mathrm{Ker}\, g$, alors g est une injection, car, si $g(x) = g(y)$, on a

$$g(x - y) = g(x) - g(y) = 0,$$

donc $x - y \in \mathrm{Ker}\, g$, d'où $x - y = 0$, soit $x = y$. Cette propriété, qui permet en fait de « translater » toutes les autres, est bien particulière aux applications linéaires.

Pour étudier si l'on a une surjection, on doit considérer l'ensemble des points de F qui sont images d'un élément de E au moins. On introduit alors l'ensemble des points-images, soit l'image de g, que l'on note $\mathrm{Im}\, g$, qui est un sous-espace vectoriel de F. Ces ensembles, lorsque E est de dimension finie, sont liés par la relation fondamentale :

$$\dim_K E = \dim_K \mathrm{Ker}\, g + \dim_K \mathrm{Im}\, g.$$

En particulier, si g est une application linéaire de E dans lui-même (ou endomorphisme), on obtient le résultat :

$$g \text{ injective} \Leftrightarrow g \text{ surjective} \Leftrightarrow g \text{ bijective}$$

que l'on peut même étendre aux applications de E dans F, pourvu que les deux espaces vectoriels soient de dimensions finies.

Lorsque F = K, on dit que l'on a une forme linéaire ; ceci a bien un sens puisque K est un espace vectoriel sur lui-même. Dans ce cas, le sous-espace $\mathrm{Im}\, g$ est inclus dans K ; il est donc de dimension 0 ou 1, et le noyau de g, si E est de dimension finie, est donc de dimension n ou $n - 1$. Dans le premier cas, la forme linéaire est la forme nulle (tout vecteur a pour image 0) ; et dans le second cas, on dit que le noyau est un hyperplan.

L'ensemble des applications linéaires de E dans F est un espace vectoriel sur le même corps, lorsqu'on y définit les opérations d'addition et de produit par un scalaire usuelles ; on le note $\mathcal{L}(E, F)$. Lorsque E = F, la composition des applications lui donne une structure d'algèbre, et on le note $\mathcal{L}(E)$. Enfin, si F = K, on obtient l'ensemble des formes linéaires sur E, que l'on nomme *dual* de E, noté E*.

On va regarder maintenant comment l'on peut caractériser un élément quelconque de $\mathcal{L}(E, F)$, et l'on s'y intéressera plus spécialement pour les espaces vectoriels E et F de dimension finie.

Une base $B = \{e_i\}_{i \in I}$ de E étant donnée, tout élément $x \in E$ possède une décomposition unique $x = \sum_{i \in I} \alpha_i e_i$, selon cette base (et où seul un nombre fini de termes ne sont pas nuls). Soit $g \in \mathcal{L}(E, F)$, alors, en raison des conditions de linéarité de g, on a :

$$\varphi(x) = \sum_{i \in I} \alpha_i\, \varphi(e_i)$$

La connaissance de x équivaut à celle de la suite $(\alpha_i)_{i \in I}$, donc $g(x)$ sera entièrement connu par la donnée de tous les éléments $g(e_i)$, images des vecteurs de la base B. Ceci revient à dire qu'une application linéaire est déterminée totalement par les images des vecteurs d'une base de l'espace de départ. De cette idée découle toute la théorie des matrices, introduite par Sylvester, et dont les développements sont immenses.

Matrices — Calcul matriciel

L'utilisation des tableaux rectangulaires de nombres est connue depuis très longtemps, et de nos jours, il n'est pratiquement pas de domaine (où les nombres jouent un rôle) qui ne les utilise. Toutefois, ce n'est qu'au XIXe siècle que le mathématicien anglais Arthur Cayley précisa la

notion de matrice comme étant une forme de représentation, disons un « modèle », pour une transformation linéaire. Il en déduisit un exposé clair du calcul matriciel, lequel a pour but essentiel de simplifier toutes les écritures ayant trait aux opérations sur des applications linéaires. Cette justification — sur des bases mathématiques claires — de l'utilisation des tableaux de nombres a permis tous les plus fructueux développements que l'on connaît actuellement.

Soit deux espaces vectoriels réels ou complexes E et F, de dimensions finies n et p (ce que nous supposerons dans tout ce paragraphe) ; une base $B_E = \{e_1, e_2, ..., e_n\}$ de E, une base $B_F = \{f_1, f_2, ..., f_p\}$ de F et $g \in \mathcal{L}$ (E, F). Alors, pour tout élément $x = \sum_{j=1}^{j=n} \alpha_j e_j$, on a $\varphi(x) = \sum_{j=1}^{j=n} \alpha_j\, g(e_j)$.

Puisque $g(e_j) \in F$ pour toutes les valeurs de j, chacun de ces vecteurs possède une décomposition dans la base B :

$$g(e_j) = \sum_{i=1}^{i=p} \beta_{ij} f_i$$

Cette fois, l'écriture comporte deux indices pour montrer que, pour chaque élément $g(e_j)$, il existe une suite de p coefficients. On dispose par conséquent de $n \times p$ nombres réels ou complexes β_{ij} qui définissent entièrement tous les vecteurs $g(e_1)$, $g(e_2)$, ..., $g(e_n)$, donc l'application g elle-même. Très exactement, β_{ij} est la i-ième composante du vecteur $g(e_j)$, c'est-à-dire sa coordonnée selon le vecteur f_i de la base B_F. Par convention, en rangeant colonne par colonne les composantes des vecteurs $g(e_j)$, on obtient la *matrice* de l'application g relativement aux bases B_E et B_F :

$$\begin{pmatrix} \beta_{11} & \beta_{12} \dots \beta_{1j} \dots \beta_{1n} \\ \beta_{21} & \beta_{22} \dots \beta_{2j} \dots \beta_{2n} \\ \vdots & \vdots \quad\;\; \vdots \\ \beta_{i1} & \beta_{i2} \dots \beta_{ij} \dots \beta_{in} \\ \vdots & \vdots \quad\;\; \vdots \quad\;\; \vdots \\ \beta_{p1} & \beta_{p2} \dots \beta_{pj} \dots \beta_{pn} \end{pmatrix}$$

On dit que l'on a une matrice (p, n), ou encore à p lignes et n colonnes, et l'on note, en abrégé, cette matrice :

$$(\beta_{ij})_{\substack{i=1,2,...,p \\ j=1,2,...,n}}$$

(n et p sont les dimensions de la matrice).

Voici quelques matrices remarquables dans le cas E = F (les matrices sont alors des matrices carrées), on identifiera cet espace vectoriel, grâce à l'isomorphisme vu plus haut, à $\mathbb{R}^n$ (ou $\mathbb{C}^n$ dans le cas d'un espace vectoriel complexe) muni de sa base canonique.

La matrice de $g : x \to 0$ est donc $(0)_{\substack{i=1,2,...,n \\ j=1,2,...,n}}$ (matrice nulle).

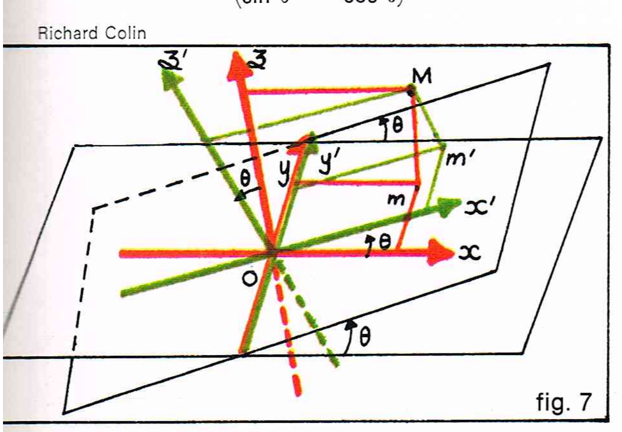
La matrice de $g : x \to x$ est alors la matrice identité où seuls ne sont pas nuls — et valent 1 — les éléments de la diagonale, c'est-à-dire ceux pour lesquels les deux indices sont égaux.

La matrice de $g : x \to \alpha x$, où α est un scalaire, est $\begin{pmatrix} \alpha & 0 \\ 0 & \alpha \end{pmatrix}$.

La rotation d'angle θ autour de l'origine, application de $\mathbb{R}^2$ dans $\mathbb{R}^2$, est représentée par la matrice :

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

Richard Colin



La rotation d'angle θ, autour de l'axe Oy (*fig.* 7), application de $\mathbb{R}^3$ dans $\mathbb{R}^2$, admet pour matrice :

$$\begin{pmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{pmatrix}$$

Soit $A = (a_{ij})_{\substack{i=1...p \\ j=1...n}}$ une matrice (p, n) et

$$X = \sum_{i=1}^{n} x_i e_i \in E.$$

On définit le produit AX pour que le résultat soit l'image de X par l'application linéaire : $g : E \to F$ dont A est la matrice relativement aux bases canoniques. Puisque :

$$g(X) = \sum_{i=1}^{p} x_i\, g(e_i), \quad \text{on a } g(X) = \sum_{i=1}^{n} x_i \sum_{j=1}^{p} a_{ji} f_j$$

d'après la définition d'une matrice. On a donc un vecteur à p composantes dont la j-ième (le facteur de f) vaut exactement $\sum_{i=1}^{n} x_i a_{ji}$. Or la j-ième ligne de A s'écrit : $(a_{1j}, a_{2j}, ..., a_{nj})$, et si l'on définit le produit de cette ligne par le vecteur-colonne des x_i comme étant égal à :

$$x_1 a_{1j} + x_2 a_{2j} + \dots + x_n a_{nj}$$

(somme des produits de termes se correspondant), alors le produit de la matrice A (matrice (p, n)) par le vecteur-colonne X (ou matrice $(n, 1)$) est obtenu en faisant le produit de chaque ligne de matrice A par le vecteur X comme on vient de la définir (et qui n'est que la généralisation de la notion classique du produit scalaire de deux vecteurs, mais dans $\mathbb{R}^n$), et en rangeant les résultats obtenus en un vecteur-colonne.

$$(a_{j1}\, a_{j2}\, \dots\, a_{jn}) \cdot \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{pmatrix} = a_{j1} x_1 + a_{j2} x_2 + \dots + a_{jn} x_n$$

$$\begin{pmatrix} a_{11} & \dots & a_{1n} \\ a_{21} & \dots & a_{2n} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ a_{p1} & \dots & a_{pn} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{j=n} a_{1j} x_j \\ \sum_{j=1}^{j=n} a_{2j} x_j \\ \cdot \\ \cdot \\ \cdot \\ \sum_{j=1}^{j=n} a_{pj} x_j \end{pmatrix}$$

Opérations usuelles

De façon générale, le calcul matriciel s'obtient en « décalquant » sur les matrices telles qu'elles sont définies, les opérations usuelles sur les applications, et leurs propriétés.

Posons donc

$$A = (a_{ij})_{\substack{i=1,2,...p \\ j=1,2,...n}} \quad \text{et} \quad B = (b_{ij})_{\substack{i=1,2,...,p \\ j=1,2,...,n}}$$

deux matrices (p, n).

On dit qu'elles sont égales si, pour toutes les valeurs de i et de j, on a : $a_{ij} = b_{ij}$. En effet, deux applications linéaires sont égales si les images des vecteurs de base coïncident.

La somme des deux matrices A et B est la matrice C, matrice (n, p), obtenue en faisant la somme des termes correspondants de A et B : $c_{ij} = a_{ij} + b_{ij}$ pour toutes valeurs de i et de j ; si A représente l'application g_1 et si B représente l'application g_2 par rapport à deux bases données, alors C = A + B représente $g_1 + g_2$ dans ces mêmes bases.

Soit maintenant $k \in K$, un élément du corps de base (réel ou complexe). Alors, kA est la représentation de l'application kg si l'on pose :

$$kA = (ka_{ij})_{\substack{i=1,2,...,p \\ j=1,2,...,n}}$$

◀ **Figure 7 : *rotation d'angle* θ *autour de l'axe* Oy.**

Ces deux opérations donnent à l'ensemble des matrices (p, n), à coefficients dans le corps K, une structure d'espace vectoriel sur K. Cet espace se désigne par $\mathcal{M}_{p,n}(K)$; il est isomorphe à $\mathcal{L}(E, F)$, et l'on a donc :

$$\dim_K \mathcal{L}(E, F) = \dim_K \mathcal{M}_{p,n}(K) = n \times p.$$

Toutefois il faut noter que cet isomorphisme dépend étroitement des bases choisies de E et de F ; en ce sens, on dit qu'il n'est pas canonique. Cependant c'est un moyen bien commode pour identifier les deux ensembles. Une application linéaire ne peut donc être associée à une matrice que par l'intermédiaire d'une base de chacun des espaces vectoriels E et F. On verra justement plus loin l'effet d'un changement de base sur une matrice.

Le produit matriciel

Soit E, F, G trois espaces vectoriels sur un même corps K. Soit $g_1 \in \mathcal{L}(E, F)$ et $g_2 \in \mathcal{L}(F, G)$, on sait que $g_2 \circ g_1(x) = g_2[g_1(x)]$ pour $x \in E$; c'est l'opération de composition des applications. Le produit matriciel est une opération entre matrices de telle sorte que B · A (si B et A sont respectivement les matrices de g_2 et g_1 relativement à des bases données) soit la matrice associée à l'application $g_2 \circ g_1$ (par rapport aux bases déjà choisies). Les conditions dans lesquelles on peut réaliser cette opération sont définies par le fait que F, espace d'arrivée pour g_1, est aussi l'espace de départ pour g_2 ; la dimension de F devra donc être aussi bien le nombre de colonnes de B que le nombre de lignes de A. D'autre part, si $n = \dim_K E$, $p = \dim_K F$ et $q = \dim_K G$, alors A est une matrice (p, n), B une matrice (q, p) et C (matrice de $g_2 \circ g_1$) une matrice (q, n).

D'où la règle suivante : le produit de deux matrices B et A n'est réalisable que si le nombre de colonnes de B est égal au nombre de lignes de A. Dans ce cas, le produit d'une matrice (q, p) par une matrice (p, n) est une matrice (q, n).

Sans chercher à montrer rigoureusement — ce qui nécessiterait quelques calculs lourds — la méthode du produit matriciel, on la justifie en notant qu'une matrice est une collection de vecteurs-colonnes. Il est donc logique d'avoir à multiplier la première matrice par chaque colonne de la seconde ; ceci amène à formuler que le produit de B par la première colonne de A donne la première colonne de C, et ainsi de suite jusqu'à la dernière colonne de A qui donne — par produit avec B — la dernière colonne de C. On peut donc énoncer la formule du produit matriciel :

$$c_{ij} = \sum_{k=1}^{k=p} b_{ik} a_{kj} = b_{i1} a_{1j} + b_{i2} a_{2j} + \dots + b_{ip} a_{pj}$$

Autrement dit, le terme situé à la i-ième ligne et j-ième colonne de C = B × A est le produit de la i-ième ligne de B par la j-ième colonne de A, au sens déjà défini plus haut.

Cette règle est schématisée ci-dessous :

$$\begin{pmatrix} b_{i1}\, b_{i2} \,\dots\dots\dots\dots\, b_{ip} \end{pmatrix} \dots \begin{pmatrix} & \vdots \\ \dots & c_{ij} \\ & \vdots \end{pmatrix}$$
$$\begin{pmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{pj} \end{pmatrix}$$

où $C_{ij} = (b_{i1}\, b_{i2} \dots b_{ip}) \begin{pmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{pj} \end{pmatrix}$

$$= b_{i1} a_{1j} + b_{i2} a_{2j} + \dots + b_{ip} a_{pj}$$

Les propriétés suivantes, connues pour les applications linéaires, ne surprendront donc pas.

Propriété 1. Le produit matriciel n'est pas commutatif. Il n'est d'ailleurs pas toujours possible de réaliser A × B lorsqu'on peut calculer B × A. Cela ne peut se faire que si le nombre de colonnes de A est égal au nombre de lignes de A *et* si le nombre de colonnes de B est égal au nombre de lignes de A. Donc, si A est une matrice (p, n), il faut que B soit une matrice (n, p) ; alors AB est une matrice (p, p) et BA une matrice (n, n), qui ne peuvent être égales si $n = p$. Mais, même si $n = p$, de toutes façons la commutativité n'est pas règle mais exception.

Propriété 2. Soit $I_n = (\delta_{ij})_{\substack{i=1,2,\dots,n \\ j=1,2,\dots,n}}$ la matrice identité d'ordre n.
Alors, pour toute matrice A d'ordre n, on a :

$$A \times I_n = I_n \times A = A.$$

En effet, I_n correspond à la transformation identité.

Propriété 3. Le produit matriciel est associatif :

$$(AB)\,C = A\,(BA) = ABC.$$

Propriété 4. Le produit matriciel est distributif par rapport à la somme :

$$A \cdot (B + C) = A \cdot B + A \cdot C$$
$$(A + B) \cdot C = A \cdot C + B \cdot C$$

et, de plus, pour tout élément $k \in K$, on a :

$$(kB) \cdot C = k\,(B \cdot C) = B \cdot (kC).$$

Propriété 5. Soit $0_{p,n} = (0)_{\substack{i=1,2,\dots,p \\ j=1,2,\dots,n}}$ la matrice nulle de dimensions (p, n).
Alors,
pour toute matrice A, de dimension (n, q) :

$$0_{p,n} \times A = 0_{p,q}$$

et pour toute matrice B, de dimension (q, p) :

$$B \times 0_{p,n} = 0_{q,n}$$

Avant de formuler quelques exemples, notons le cas particulier de l'ensemble des matrices carrées d'ordre n, soit $\mathcal{M}_n(K)$. C'est un espace vectoriel de dimension n^2. Sur cet ensemble, on peut définir une seconde loi de composition interne : le produit matriciel (de même que sur l'ensemble $\mathcal{L}(E)$ on peut définir le produit de composition des applications). Chacun des deux ensembles $\mathcal{L}(E)$ et $\mathcal{M}_n(K)$ est alors muni d'une structure d'algèbre non commutative.

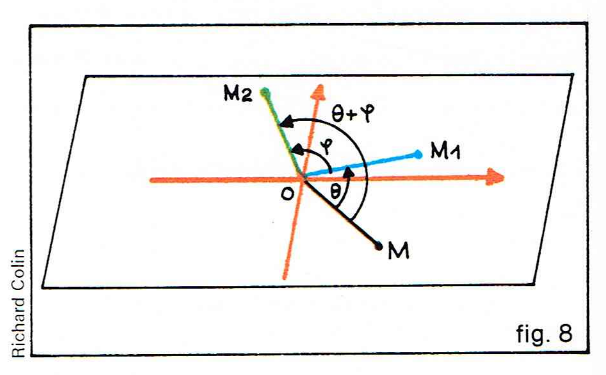
La rotation d'angle θ autour de l'origine dans $\mathbb{R}^2$ est représentée par la matrice, relativement à la base canonique,

$$M\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

A l'aide des formules de trigonométrie, on peut voir que :

$$M_\theta \times M_\varphi = M_\varphi \times M_\theta = M_{\theta+\varphi}$$

ce qui s'explique clairement sur la figure 8.



▶ Figure 8 : *rotation d'angle θ autour de l'origine O (dans $\mathbb{R}^2$) composée avec une rotation d'angle φ autour de O :*
$(\overrightarrow{OM}, \overrightarrow{OM_1}) = \theta$
$(\overrightarrow{OM_1}, \overrightarrow{OM_2}) = \varphi$
$(\overrightarrow{OM}, \overrightarrow{OM_2}) = \theta + \varphi$.

On en déduit que $M_\theta \times M_{-\theta} = M_{-\theta} \times M_{-\theta} = I_2$. Dans cet exemple, il est clair que, pour toute matrice M_θ, il existe une matrice (soit $M_{-\theta}$) qui, multipliée par M_θ, donne l'identité : on dit que $M_{-\theta}$ est la matrice inverse de M_θ. Toute matrice $A \in \mathcal{M}_n(K)$ ne possède pas cette propriété : en effet, ceci revient à étudier si l'application linéaire f dont A est la matrice — relativement à une base donnée — possède une application réciproque ; ceci n'a lieu que si f est bijective, ou encore injective puisque l'on est en dimension finie. La propriété suivante permettra de caractériser les matrices possédant une inverse.

L'application f est injective si, et seulement si, les vecteurs $f(e_1), f(e_2), \dots, f(e_n)$ forment un système libre dès que $\{e_1, e_2, \dots, e_n\}$ est une base de E.

Soit en effet $S = \{f(e_1), f(e_2), \dots, f(e_n)\}$, qui est libre si $\sum_{i=1}^{i=n} \alpha_i f(e_i) = 0$ implique que $\alpha_i = 0$ pour toutes les

valeurs de i. Mais, par linéarité, on a :

$$\sum_{i=1}^{n} \alpha_i f(e_i) = \sum_{i=1}^{n} f(\alpha_i e_i) = f\left(\sum_{i=1}^{n} \alpha_i e_i\right)$$

et on sait que :

$$x = \sum_{i=1}^{n} \alpha_i e_1 \in E$$

Donc, dire que S est libre, c'est dire que $f(X) = 0$ implique $X = 0$, donc que f est injective.

Réciproquement, si f est injective, pour tout $X \in E$, donc de la forme $X = \sum_{i=1}^{n} a_i e_i$, on a $f(X) = 0$ seulement si $X = 0$.

D'autre part, la linéarité de f donne :

$$f\left(\sum_{i=1}^{n} a_i e_i\right) = \sum_{i=1}^{n} a_i f(e_i).$$

On a donc les implications :

$$f(X) = f\left(\sum_{i=1}^{n} a_i e_i\right) = \sum_{i=1}^{n} a_i f(e_i) = 0 \Rightarrow$$

$$\sum_{i=1}^{n} a_i e_i = 0 \Rightarrow a_i = 0$$

pour toute valeur de i ; la seconde implication étant la transcription du fait que $(e_i)_{i=1,2,...,n}$ est une base, donc un système libre. Cette propriété donne le résultat fondamental qui suit.

Théorème. La matrice $A \in \mathcal{M}_n(K)$ est inversible (possède une inverse) si, et seulement si, les vecteurs-colonnes de A forment un système libre.

On peut remarquer que pour une matrice (2, 2) qui s'écrit $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$, les vecteurs-colonnes sont linéairement indépendants si :

$$\alpha_1 \begin{pmatrix} a \\ c \end{pmatrix} + \alpha_2 \begin{pmatrix} b \\ d \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow \alpha_1 = \alpha_2 = 0$$

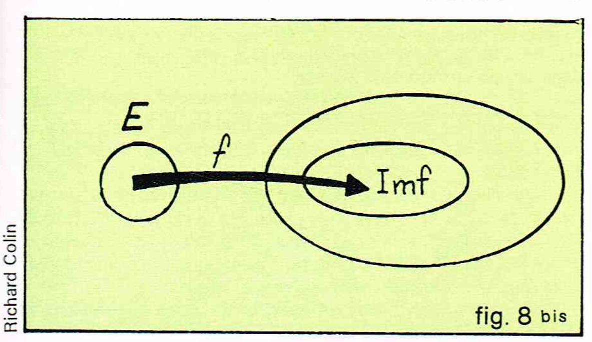
soit encore :

$$\begin{pmatrix} \alpha_1 a + \alpha_2 b \\ \alpha_1 c + \alpha_2 d \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow \alpha_1 = \alpha_2 = 0.$$

Or les équations simultanées : $\begin{cases} \alpha_1 a + \alpha_2 b = 0 \\ \alpha_1 c + \alpha_2 d = 0 \end{cases}$ ont la solution $\alpha_1 = \alpha_2 = 0$, unique, sauf si $\frac{a}{c} = \frac{b}{d}$, auquel cas α_2 peut être choisi quelconque et $\alpha_1 = -\frac{b}{a}\alpha_2$ dès que $a \neq 0$. On reconnaît là la notion de dépendance linéaire exprimée sous forme de proportionnalité.

Rang d'une application linéaire.

Si la matrice A n'est pas inversible, l'application linéaire f à laquelle elle est associée n'est pas bijective, donc pas surjective. On pense donc alors à caractériser l'ensemble Im f, sous-espace vectoriel de F ($f : E \to F$) tel que $\dim_K \text{Im} f < \dim_K F$ (sinon Im f = F et f serait surjective) par sa dimension. Ce nombre, qui est encore le nombre maximal de vecteurs-colonnes de A linéairement indépendants, s'appelle le rang de f et se note $rg(f)$ [*fig.* 8 bis].

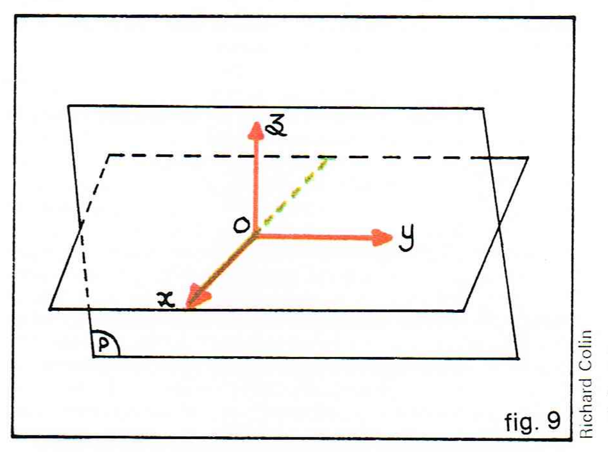


◀ **Figure 8 bis :** ***rang d'une application linéaire.***

On peut montrer que ce nombre est caractéristique de f, et ne dépend pas des bases choisies pour la matrice A. Il joue un rôle fondamental dans la théorie des systèmes d'équations linéaires. Plus généralement encore, on dit que $f \in \mathcal{L}(E, F)$ est de rang fini si le sous-espace vectoriel Im f est de dimension finie (ce, pour le cas où E et F sont de dimensions quelconques). Bien évidemment, pour une matrice carrée A inversible, on a rg A = dim E.

Pour illustrer ces notions, on considère l'exemple suivant :

soit $E = \mathbb{R}^3$ et $F = \mathbb{R}^2$ (géométriquement représentés par l'espace à 3 dimensions, et le plan « horizontal » des deux premières coordonnées) et l'application $f : \mathbb{R}^3 \to \mathbb{R}^2$ définie en tout point par : $f(x, y, z) = (x, 0)$. Cette application (*fig.* 9) n'est autre que la projection sur le premier axe de $\mathbb{R}^2$; elle n'est pas surjective puisque les images ont concentrées seulement sur le premier axe du plan. L'ensemble Im f est donc très exactement cette droite ; il est donc de dimension 1. On peut noter que l'ensemble Ker f est défini par les points $(0, y, z)$ où y et z sont quelconques, car $f(0, y, z) = (0, 0)$. Cet ensemble est donc le plan yOz défini par le second et le troisième axes de $\mathbb{R}^3$. On aurait pu aussi directement voir que la matrice de f relativement aux bases canoniques de $\mathbb{R}^2$ et $\mathbb{R}^3$ s'écrit $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$ et que le seul système libre parmi les vecteurs-colonnes est formé par le premier.



◀ **Figure 9 :** ***pour la projection* p :** **(x, y, z) → (x, O, O)** ***de*** $\mathbb{R}^3$ ***sur l'axe* Ox*, le plan* (P) *défini par* Oy *et* Oz *représente* Ker (p), *et l'axe* Ox *représente* Im (p).**

Changement de base

Pour finir cette brève présentation des outils du calcul matriciel, on va étudier l'effet d'un changement de base sur une matrice carrée (on se bornera à ce seul cas).

Soit donc $\{e_1, e_2, ..., e_n\}$ une base de E, et $\{v_1, v_2, ..., v_n\}$ une nouvelle base de E dont chacun des vecteurs s'exprime en fonction de l'ancienne :

$$v_1 = \sum_{j=1}^{n} a_{1j} e_j, \quad v_2 = \sum_{j=1}^{n} a_{2j} e_j, \ ..., \quad v_n = \sum_{j=1}^{n} a_{nj} e_j$$

(les indices doubles nous permettent de garder la même lettre de référence pour les composantes tout en les différenciant pour chaque décomposition). La matrice :

$$A = (a_{ij})_{\substack{i=1,...,n \\ j=1,...,n}}$$

est la représentation de l'application identité par rapport, respectivement, aux bases $(v_i)_{i=1,...,n}$ et $(e_i)_{i=1,...,n}$ (le lecteur peut vérifier que l'image de v_i, donc v_i lui-même, s'écrit $a_{i1} e_1 + a_{i2} e_2 + ... + a_{in} e_n$, d'où l'expression de la matrice A).

Prenons alors un vecteur $X \in E$; dans la première base, il se décompose en $\sum_{i=1}^{n} x_i e_i$ et, dans la seconde, en $\sum_{i=1}^{n} p_i v_i$.

L'expression de v_i dans la première base, soit $\sum_{j=1}^{n} a_{ij} e_j$, donne donc :

$$X = \sum_{i=1}^{n} \sum_{j=1}^{n} p_i a_{ij} e_j = \sum_{j=1}^{n} \left(\sum_{i=1}^{n} p_i a_{ij}\right) e_j.$$

Appelons alors $X_{(ei)}$ et $X_{(vi)}$ les vecteurs-colonnes

représentant X dans les bases $(e_i)_{i=1,\ldots,n}$ et $(v_i)_{i=1,\ldots,n}$ respectivement ; la relation :

$X_{(ei)} = A \times X_{(vi)}$ **(1)** traduit le résultat ainsi obtenu ; dans cette égalité, A est donc la matrice où le k-ième vecteur-colonne est formé des composantes k dans la base $(e_i)_{i=1,\ldots,n}$, du vecteur V_k.

Puisque A est une matrice inversible (car l'application associée est bijective), on peut multiplier, à gauche, chaque membre de l'égalité **(1)** par la matrice A^{-1}, et on obtient $A^{-1} X_{(ei)} = X_{(vi)}$ (puisque $AA^{-1} = I_n$).

Soit maintenant $f \in \mathcal{L}$ (E), M sa matrice relativement à la base $(e_i)_{i=1,\ldots,n}$ et M' sa matrice relativement à la base $(v_i)_{i=1,\ldots,n}$. Pour tout élément $x \in E$, soit $y = f(x)$, ce qui s'écrit :

$$Y_{(ei)} = MX_{(ei)} \quad \text{et} \quad Y_{(vi)} = M'X_{(vi)}.$$

De plus :

$$X_{(ei)} = AX_{(vi)} \quad \text{et} \quad Y_{(ei)} = AY_{(vi)} ;$$

donc $\quad Y_{(ei)} = MAX_{(vi)} = AY_{(vi)}$

ce qui donne encore :

$$Y_{(vi)} = A^{-1} MAX_{(vi)} ; \quad \text{mais } Y_{(vi)} = M'X_{(vi)}$$

donc, par comparaison, on obtient la relation fondamentale :

$$M' = A^{-1} \cdot M \cdot A$$

On montrera plus loin des applications de cet important résultat. Ajoutons simplement que deux matrices carrées A et B sont dites *équivalentes*, si elles représentent le même endomorphisme mais dans deux bases différentes ; elles sont donc telles qu'il existe une matrice Q inversible vérifiant $B = Q^{-1} \cdot A \cdot Q$. Il s'agit là d'une relation d'équivalence entre matrices carrées de même ordre.

La dualité

Si E est un espace vectoriel sur un corps K, l'ensemble des applications linéaires de E dans K, ou encore formes linéaires sur E, est un espace vectoriel sur K pour les deux lois classiques ; on l'appelle *espace dual* de E, et on le note E*. Cette idée peut se prolonger, et on peut ainsi définir l'ensemble des applications linéaires de E* dans K (formes linéaires sur E*) que l'on appelle *bidual* de E, noté E**. Des liens étroits peuvent être tracés entre les espaces E, E* et E** ; et, malgré son apparence abstraite (tel que nous venons de le poser), ce problème est né sur des idées très concrètes de géométrie, par l'étude des transformations par polaires réciproques (due à Poncelet) et des formes quadratiques (Möbius, Chasles, etc.).

On a vu que, si $\dim_K E = n$ et $\dim_K F = p$, alors $\dim_K \mathcal{L}(E, F) = np$. On en déduit que, si E est de dimension finie n, alors E* est aussi de dimension finie n (toutefois, si E est de dimension infinie, E* est aussi de dimension infinie, mais il n'existe plus d'isomorphisme entre eux). L'isomorphie de E et E* va être définie par une base de E*. Soit $(e_i)_{i=1,\ldots,n}$ une base de E. Tout élément $x \in E$ s'écrit $x = \sum_{i=1}^{n} x_i e_i$, et cette décomposition est unique. Pour chaque valeur de l'indice i, l'application $e^i : x \to x_i$ est une forme linéaire sur E. On a ainsi n éléments du dual E*, soit $e^1, e^2, \ldots, e^n$, qui forment un système libre — ce qu'on montre aisément — donc une base de E*, qu'on nomme *base duale* de la base $(e_i)_{i=1,\ldots,n}$. Les formes linéaires e^i peuvent être définies, de façon concise, par :

$$e^i(e_j) = 0 \quad \text{si} \quad i \neq j$$
$$e^i(e_i) = 1$$

que l'on peut encore simplifier par l'usage du symbole de Kronecker δ_{ij} défini par :

$$\delta_{ij} = \begin{cases} 1 & \text{si} \quad i = j \\ 0 & \text{si} \quad i \neq j \end{cases}$$

On a alors :

$$e^i(e_j) = \delta_{ij}$$

L'isomorphisme [à la base $(e_i)_{i=1,\ldots,n}$, on fait correspondre la base $(e^i)_{i=1,\ldots,n}$] ainsi construit n'est pas canonique puisqu'il dépend de la base choisie dans l'espace E.

Formes bilinéaires

Soit E et F deux espaces vectoriels sur le même corps de base K. L'ensemble produit E × F peut être muni d'une structure d'espace vectoriel sur K par les 2 lois :

$$(a, \alpha) + (b, \beta) = (a + b, \alpha + \beta) \quad \text{si } a \in E, b \in E, \alpha \in F, \beta \in F$$
$$\lambda (a, \alpha) = (\lambda a, \lambda \alpha) \quad \text{si} \quad \lambda \in K, a \in E \text{ et } \alpha \in F.$$

Les axiomes de structure se vérifient aisément sur *l'espace vectoriel produit* E × F, et l'on montre que si E et F sont de dimensions finies, alors

$$\dim_K (E \times F) = \dim_K E + \dim_K F.$$

Appelons *forme bilinéaire* définie sur E × F une application de E × F dans K qui soit linéaire par rapport à chacun de ses arguments. Une telle application

$$f = E \times F \to K$$

doit donc vérifier :

$$f(a + kb, \alpha) = f(a, \alpha) + kf(b, \alpha)$$
$$f(a, \alpha + k\beta) = f(a, \alpha) + kf(a, \beta).$$

Une forme bilinéaire bien connue est celle dite produit scalaire : si E et F sont de même dimension finie n sur K, soit $x \in E$ décomposé sous la forme $(x_1, x_2, \ldots, x_n)$ relativement à une base donnée de E, et soit $y \in F$ décomposé en $(y_1, y_2, \ldots, y_n)$ par rapport à une base donnée de F ; alors $(x, y) \to x_1 y_1 + x_2 y_2 + \ldots + x_n y_n = \sum_{i=1}^{n} x_i y_i$ est une forme bilinéaire sur E × F, appelée *produit scalaire* de x et y. Si $E = F = \mathbb{R}^2$ ou si $E = F = \mathbb{R}^3$, on retrouve la notion analytique du produit scalaire dans le plan ou dans l'espace. On définit une forme bilinéaire sur E × E* (dite produit scalaire sur E × E*) par :

$$(x, f) \to \langle x, f \rangle = f(x) \quad \text{si} \quad f \in E^* \quad \text{et} \quad x \in E.$$

La notation $\langle x, f \rangle$ est plus spécialement utilisée dans les questions relatives à l'espace dual ; on dit que l'on a des crochets de dualité. Cette application est canonique puisque le résultat ne dépend pas de bases choisies entre E et E*. En fixant $x \in E$, et en faisant varier $f \in E^*$, on obtient une forme linéaire sur E*, donc un élément — lié à x — du bidual E**, et que l'on notera $\hat{x}$; $\hat{x} : f \to \langle x, f \rangle$. L'application de E dans E** ainsi mise en évidence, qui à tout $\hat{x} \in E$ associe $x \in E^{**}$, est linéaire ; c'est un morphisme canonique de E dans E** qui va permettre de dégager les liens de dualité.

Tout d'abord, dans le cas où l'espace E est de dimension finie n, on a encore

$$\dim_K E^{**} = \dim_K (E^*)^* = \dim_K E^* = \dim_K E = n.$$

De plus supposons que l'on ait deux éléments $\hat{x}$ et $\hat{y}$ égaux, alors $\langle \hat{x}, f \rangle = \langle \hat{y}, f \rangle$ pour tout $f \in E^*$, donc $\langle \hat{x} - \hat{y}, f \rangle = 0$, en prenant successivement pour f les formes $e^1, e^2, \ldots, e^n$, on obtient :

$$x_1 - y_1 = 0, \quad x_2 - y_2 = 0, \quad \ldots, \quad x_n - y_n = 0$$

donc $x = y$.

Par conséquent, le morphisme $x \to \hat{x}$ est injectif, donc bijectif, ce qui entraîne : si E est de dimension finie, les espaces E et E** sont canoniquement isomorphes.

Il est donc possible d'identifier E et E** (à un isomorphisme près), et par conséquent de considérer E comme le dual de E*. La dualité est ainsi un phénomène « symétrique » présentant une réciprocité : E et E* sont duaux l'un de l'autre. Cette propriété permet de conclure que chaque élément de E peut être considéré comme une forme linéaire sur E* (en identifiant x et $\hat{x}$ définis dans l'isomorphisme canonique).

Dans le cas où l'espace vectoriel E est de dimension infinie, il faut noter que l'homomorphisme $x \to \hat{x}$ est injectif, mais non surjectif. L'isomorphisme, mis en évidence si $\dim_K E < +\infty$ (*fig.* 10), n'en est plus un en dimension infinie.

Richard Colin

fig. 10

▶ **Figure 10 :** a, *en dimension finie,* E *et* E** *sont isomorphes ;* b, *en dimension infinie,* φ *est seulement surjective.*

Transposition.

Soit E et F deux espaces vectoriels sur le corps K, et $f \in \mathcal{L}(E, F)$. Pour tout $g \in \mathcal{L}(F^*, E^*)$ et tout $\alpha \in F^*$, $g(\alpha) \in E^*$, donc $g(\alpha) : E \to K$; d'autre part $f : E \to F$ et $\alpha : F \to K$ impliquent que $\alpha \circ f : E \to K$. On peut alors montrer qu'il existe une application $g \in \mathcal{L}(F^*, E^*)$ et une seule telle que : $g(\alpha) = \alpha \circ f$. Cette application se nomme transposée de f, ce qui s'écrit ${}^t f$, et est donc définie par :

(1) $\langle {}^t f(\alpha), x \rangle = \langle \alpha, f(x) \rangle$
pour tout $x \in E$ et pour tout $\alpha \in F^*$.

La représentation matricielle de ${}^t f$ est très simple à déterminer. Si E est de dimension n et F de dimension p, soit $(e_j)_{j=1,\dots,n}$ une base de E et $(f_i)_{i=1,\dots,p}$ une base de F; on désigne par $(e^j)_{j=1,\dots,n}$ et $(f^i)_{i=1,\dots,p}$ les bases duales. Alors si $A = (a_{ij})_{\substack{i=1,\dots,p \\ j=1,\dots,n}}$ est la matrice de f relativement aux bases choisies de E et F, l'égalité **(1)** permet de montrer que la matrice B de ${}^t f$ relativement aux bases duales est définie par $b_{ij} = a_{ji}$.

Cela revient à dire que B est une matrice (n, p) obtenue à partir de A en inversant (en transposant, ce qui explique le nom choisi pour ${}^t f$) le rôle des lignes et des colonnes.

Les principales règles de calcul en transposition sont :

— ${}^t(f + g) = {}^t f + {}^t g \quad {}^t(kf) = k\,{}^t(f) \quad {}^t(g \circ f) = {}^t f \circ {}^t g$

— Si f admet une réciproque, il en est de même pour ${}^t f$ et $({}^t f)^{-1} = {}^t(f^{-1})$

Et de même que E et E** sont isomorphes si E est de dimension finie, on a (pour des espaces E et F de dimensions finies) un isomorphisme entre $\mathcal{L}(E, F)$ et $\mathcal{L}(F^*, E^*)$ qui n'est autre que l'application $f \to {}^t f$. Dans ce cas, on a donc : ${}^t({}^t f) = f$.

Orthogonalité

Soit E et F deux espaces vectoriels sur un même corps K et $g : E \times F \to K$ une forme bilinéaire sur $E \times F$. On dit que $x \in E$ et $\alpha \in F$ sont *orthogonaux relativement à g* si $g(x, \alpha) = 0$. Dans le cas où $F = E^*$, ceci revient à dire que $\langle x, \alpha \rangle = \alpha(x) = 0$, et on dit simplement que α et x sont orthogonaux.

Une partie A de E et une partie B de F sont orthogonales relativement à g si l'on a $g(x, \alpha) = 0$ pour tout élément de $x \in E$, et tout $\alpha \in F$. Dans le cas où A est un sous-espace vectoriel de E, l'ensemble des éléments de F orthogonaux à tous les éléments de A est un sous-espace vectoriel de F, dit *orthogonal de A dans F, relativement à g*, noté $A^\perp$. De même pour l'orthogonal d'un sous-espace vectoriel B de F, qui se note $B^\perp$; lorsque $F = E^*$, on ne précise plus la forme g. Cette notion n'est que la généralisation algébrique de la notion classique — en géométrie — d'orthogonalité.

Lorsque l'espace vectoriel E est de dimension finie n, si le sous-espace $A \subset E$ est de dimension p, l'orthogonal $A^\perp$ est de dimension $n - p$; dans ce cas (dimension finie), on a encore « symétrie » entre F (dans E) et $F^\perp$ (dans E*) puisque $(F^\perp)^\perp = F$, car alors $(F^\perp)^\perp$ est un sous-espace vectoriel de E** donc de E, de dimension $n - (n - p) = p$.

Les notions de transposition et d'orthogonalité (en dimension finie) sont liées par le résultat :

$$\text{si } f \in \mathcal{L}(E, F), \quad \text{alors } (\mathrm{Im}\, f)^\perp = \mathrm{Ker}\,({}^t f).$$

Dans le cas d'espaces de dimensions finies, le lien est alors plus précis encore entre une application et sa transposée, car alors $rg\, f = rg\,({}^t f)$ que l'on montre à l'aide du résultat précédent, et qui donne beaucoup de développements dans les techniques d'optimisation et les méthodes de résolution des systèmes d'équations linéaires.

Il est certain que la dualité s'est révélée comme une des notions les plus fécondes en mathématiques.

Systèmes d'équations linéaires

Soit E et F deux espaces vectoriels sur le même corps K. On appelle *équation linéaire* une équation de la forme $f(x) = b$ où $f \in \mathcal{L}(E, F)$ et $b \in F$. Lorsque $F = K$, alors $f \in E^*$: on a une équation linéaire scalaire. De même, un ensemble fini d'équations $f_i(x) = b_i$ où $f_i \in \mathcal{L}(E, F)$ et $b_i \in F$ s'appelle un système linéaire (scalaire lorsque $F = K$). Résoudre ces équations consiste à trouver les éléments $x \in E$ qui les vérifient. Ce type d'équations se trouve dans tous les domaines, et la mise au point des techniques de recherche des solutions, œuvre de Frōbenius (1849-1917), Kronecker (1823-1891), Weierstrass (1815-1897), etc., a été l'une des grandes étapes du développement des mathématiques. Le formalisme des espaces vectoriels donne une méthode pratique de résolution, quels que soient l'espace E, l'espace F et le nombre d'équations du système, grâce au puissant outil qu'est le calcul matriciel.

Dans ce qui suit, on suppose que E est un espace vectoriel réel de dimension finie p, muni d'une base $(e_j)_{j=1,\dots,p}$ et F un espace vectoriel réel de dimension finie n, $(f_i)_{i=1,\dots,n}$ en étant une base; on se donne $f \in \mathcal{L}(E, F)$ et $b \in F$, et on cherche l'ensemble des $x \in E$ tels que $f(x) = b$.

On peut écrire :

$$x = \sum_{j=1}^{p} x_j e_j, \quad \text{donc } f(x) = \sum_{j=1}^{p} x_j f(e_j).$$

Puisque $(f_i)_{i=1,\dots,n}$ est une base de F et que $f(e_j) \in F$, on peut aussi écrire :

$$f(e_j) = \sum_{i=1}^{n} a_{ij} f_i.$$

Et l'égalité — composante par composante — entre b et $f(x)$ donne :

$$b_i = \sum_{j=1}^{n} a_{ij} x_j \quad \text{pour } i = 1, 2, \dots, n.$$

Soit donc n équations scalaires à p inconnues. La matrice $A = (a_{ij})_{\substack{i=1,\dots,n \\ j=1,\dots,p}}$ est la matrice de f relativement aux deux bases choisies, et si l'on pose X, représentation de x dans la base $(e_j)_{j=1,\dots,p}$, et B représentation de b dans la base $(f_i)_{i=1,\dots,n}$ sous forme de vecteur-colonne, on obtient la forme matricielle $AX = B$.

Quelques remarques simples permettent de cerner le problème :

— si $b \notin \mathrm{Im}\, f$, il n'y a aucune solution;
— si $b \in \mathrm{Im}\, f$, il y a au moins une solution.
* Si $b = 0$, les solutions forment l'ensemble $\mathrm{Ker}\, f$, et il y a toujours au moins la solution dite triviale $x = 0$; le système est dit homogène.
* Si $b \neq 0$ et si $z \in \mathrm{Ker}\, f$, pour toute solution x, $x + z$ est aussi solution, car
$f(x + z) = f(x) + f(z) = f(x) = b.$

Les deux ensembles $\mathrm{Im}\, f$ et $\mathrm{Ker}\, f$ déterminent totalement les solutions. Le cas particulier où $n = p$ se discute comme suit.

(1) Si l'application f est bijective, la matrice A possède une inverse A^{-1} et, de l'équation $AX = B$, on déduit $A^{-1}AX = A^{-1}B$, soit $X = A^{-1}B$.

La solution est unique, ce qui était prévisible puisque f est bijective.

En particulier, si $b = 0$, on retrouve le fait que $\mathrm{Ker}\, f = \{0\}$. On retrouve un *système* dit *de Cramer*, du nom du mathématicien suisse Cramer (1704-1752), où le système homogène associé n'admet que la solution triviale $x = 0$, et qui admet alors une solution unique. Résoudre un système de Cramer revient à trouver l'inverse de la matrice du système; les procédés pratiques d'inversion de matrices sont exposés au paragraphe *Algorithme de Gauss ou du pivot* du chap. *calcul numérique.*

(2) Lorsque l'application f n'est pas bijective, il n'est plus nécessaire de supposer *a priori* que $n = p$, car on ne cherche pas à inverser la matrice A.

L'idée directrice est alors de se ramener, en « éliminant » si besoin est des inconnues et des équations, à un système de Cramer (la notion claire de rang d'un système est due au mathématicien allemand G. F. Frōbenius). On cherche pour cela si certaines équations ne sont pas redondantes, c'est-à-dire s'il n'y a pas d'équations linéairement dépendantes (par exemple :

$$2x + y - 3z + t = 0, \quad x + 2y - 2z + 2t = 1$$

et $x - y - z - t = 1$ sont redondantes, car la première est égale à la somme des deux autres). Autrement dit, on calcule alors le rang r de la matrice A, ce qui permet d'affirmer que l'on a réellement r équations ($r \leqslant p$ et $r \leqslant n$). Sur les p inconnues, il faudra donc en supposer $p - r$ arbitraires et résoudre le système en fonction de celles-ci. Les r inconnues qui sont alors déterminées sont dites *inconnues principales*, et les $(p - r)$ autres non

principales. Dans chaque équation, on fait passer dans le second membre tout ce qui est relatif aux équations non principales, et on obtient alors un système de Cramer de r équations à r inconnues que l'on sait résoudre.

Voici un exemple :

$$\begin{aligned} 2x - y + 4z + t &= -4 \\ 3x + 2y - 3z - 5t &= 17 \\ 5x - 3y + 8z + 2t &= -10 \end{aligned}$$

C'est un système de rang 3 ; il y aura une inconnue non principale, par exemple t, et donc indétermination d'ordre 1.

On écrit alors le système :

$$\begin{aligned} 2x - y + 4z &= -4 - t \\ 3x + 2y - 3z &= 17 + 5t \\ 5x - 3y + 8z &= -10 - 2t. \end{aligned}$$

que l'on résout par la méthode du pivot.

La théorie des systèmes linéaires est un outil fondamental pour les méthodes modernes de l'économie sans oublier les applications diverses dans toute science exacte.

Déterminants

Une forme linéaire sur un espace vectoriel E (de corps de base K) est une application de E dans K vérifiant l'hypothèse de linéarité. Cette notion s'étend à celle de *forme n-linéaire,* qui est une application de l'espace vectoriel produit $E^n = E \times E \times \ldots \times E$ (n fois), dans K, linéaire par rapport à chacun des arguments (si $n = 2$, on retrouve la notion de forme bilinéaire). Il faut ici bien distinguer la linéarité par rapport à chaque variable de la linéarité simple par rapport aux éléments de E^n. Dans le premier cas, on a :

$$\begin{aligned} f(\lambda X) = f[\lambda (x_1, x_2, \ldots, x_n)] &= f(\lambda x_1, \lambda x_2, \ldots, \lambda x_n) \\ &= \lambda f(x_1, \lambda x_2, \ldots, \lambda x_n) \\ &= \lambda^2 f(x_1, x_2, \ldots, \lambda x_n) = \lambda^n f(X) ; \end{aligned}$$

tandis que dans le second : $f(\lambda X) = \lambda f(X)$, ce qui revient à dire que l'on a ici une forme linéaire sur E.

On dit qu'une forme n-linéaire sur E, f, est alternée si elle s'annule dès que deux des arguments sont égaux ; ceci équivaut au fait que la valeur de f se change en son opposé si l'on permute deux arguments :

$$f(x_1, x_2, \ldots, x_i, \ldots, x_j, \ldots, x_n) = -f(x_1, x_2, \ldots, x_j, \ldots, x_i, \ldots, x_n)$$

L'ensemble des formes n-linéaires alternées sur E forme bien évidemment un espace vectoriel sur K pour les opérations usuelles :

$$(f, g) \to f + g \quad \text{et} \quad (\lambda, f) \to \lambda f.$$

Le résultat suivant est en fait la clef de ce paragraphe.

Si $\{x_1, x_2, \ldots, x_n\}$ forme un système lié, alors $f(x_1, x_2, \ldots, x_n) = 0$ si f est une forme n-linéaire alternée sur E.

Si f est une forme n-linéaire alternée sur E et si $f(x_1, x_2, \ldots, x_n) = 0$, alors $\{x_1, x_2, \ldots, x_n\}$ est un système lié dans E.

Il définit l'idée d'une grandeur caractérisant la dépendance ou l'indépendance d'un système de vecteurs. La définition axiomatique des déterminants a été donnée par Kronecker et Weierstrass, bien que l'outil ait été connu et utilisé depuis le début du XIXe siècle, essentiellement rattaché au calcul vectoriel, très exactement au produit extérieur (ou produit vectoriel) et au produit mixte, pour les éléments de l'espace à 3 dimensions.

Définition

On appelle *déterminant* relatif à la base $(e_i)_{i=1,\ldots,n}$ du système $\{x_1, x_2, \ldots, x_n\}$ de vecteurs de E, la valeur de la forme n-linéaire alternée unique qui prend la valeur $+1$ pour les vecteurs de la base.

Chaque vecteur se décomposant dans la base donnée : $x_k = \sum_{i=1}^{i=n} \alpha_{ik} e_i$, le déterminant du système $\{x_1, x_2, \ldots, x_n\}$ est écrit :

$$\begin{vmatrix} \alpha_{11} & \ldots & \alpha_{1i} & \ldots & \alpha_{1n} \\ \alpha_{21} & \ldots & \alpha_{2i} & \ldots & \alpha_{2n} \\ \vdots & & \vdots & & \vdots \\ \alpha_{k1} & \ldots & \alpha_{ki} & \ldots & \alpha_{kn} \\ \vdots & & \vdots & & \vdots \\ \alpha_{n1} & \ldots & \alpha_{ni} & \ldots & \alpha_{nn} \end{vmatrix}$$

avec des barres droites simples (afin de ne pas confondre avec l'écriture matricielle).

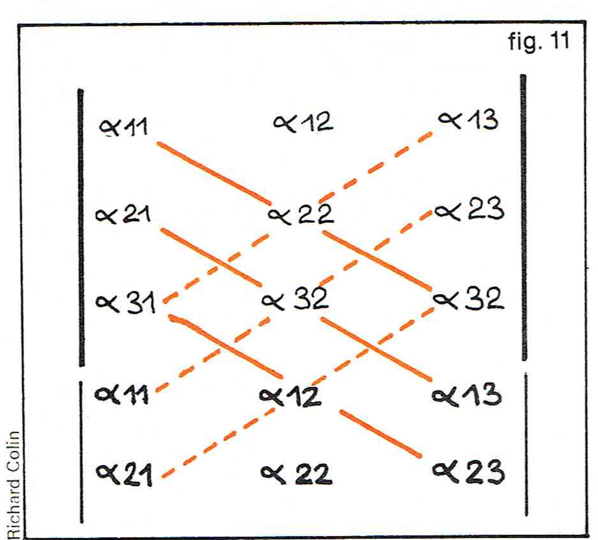
On peut considérer qu'une matrice carrée d'ordre n est formée de ses n vecteurs-colonnes, vecteurs à n composantes. On parlera donc, par abréviation, du déterminant d'une matrice au lieu du déterminant du système des vecteurs-colonnes de cette matrice.

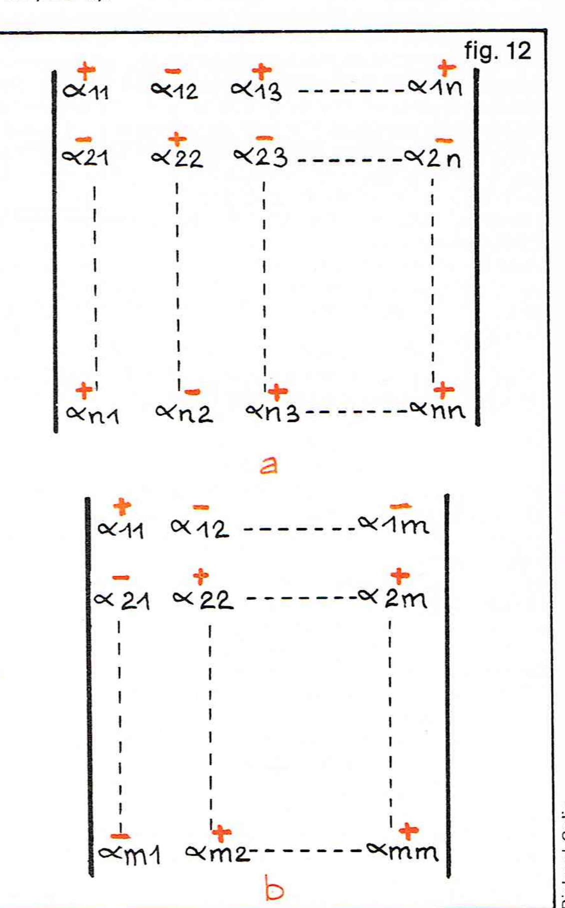
Calcul des déterminants d'ordre 1, 2 et 3

En partant du cas le plus simple, on a $|\alpha_{11}| = \alpha_{11}$ (ne pas confondre ici les barres du déterminant avec le symbole d'une valeur absolue sur $\mathbb{R}$ ou d'un module sur $\mathbb{C}$) ; puis à l'ordre 2 :

$$\begin{vmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{vmatrix} = \alpha_{11}\alpha_{22} - \alpha_{12}\alpha_{21}$$

(produit des « extrêmes » diminué du produit des « moyens »).





▶ *A gauche,* **figure 11 :** ***développement d'un déterminant d'ordre 3 par la règle de Sarrus.*** *A droite,* **figure 12 :** ***deux exemples d'affectation de signes pour le calcul d'un déterminant :*** **a,** ***d'ordre impair ;*** **b,** ***d'ordre pair.***

Enfin, à l'ordre 3, on a la règle dite *règle de Sarrus* : on reprend sous le déterminant les 2 premières lignes pour pouvoir tracer (*fig.* 11) les trois « diagonales principales » partant des éléments de la première colonne, et les trois « diagonales non principales » partant des éléments de la dernière colonne. Sur chaque « diagonale », on fait le produit des trois éléments; on affecte les résultats du signe + pour les diagonales principales, et du signe — pour les autres, et l'on obtient le nombre :

$$\alpha_{11}\alpha_{22}\alpha_{33} + \alpha_{12}\alpha_{23}\alpha_{31} + \alpha_{13}\alpha_{21}\alpha_{32}) - (\alpha_{31}\alpha_{22}\alpha_{13} + \alpha_{32}\alpha_{23}\alpha_{11} + \alpha_{33}\alpha_{21}\alpha_{12})$$

(on peut constater que c'est la généralisation du calcul à l'ordre 2).

Méthode générale de calcul

Au-delà de l'ordre 3, l'idée du calcul est de procéder par récurrence; la méthode est dite : « *développement par rapport à une ligne ou une colonne* ».

A tout élément α_{ij} du déterminant, on associe :

(1) un signe, + ou — selon sa position, égal à celui de $(-1)^{i+j}$; il découle donc de ceci que deux éléments voisins en ligne ou en colonne, mais pas en diagonale, se voient affectés de signes opposés;

(2) un déterminant d'ordre $n-1$, appelé *mineur* de α_{ij}, obtenu en supprimant dans le déterminant initial la ligne et la colonne sur lesquelles se trouve α_{ij};

(3) son *cofacteur*, égal au produit des deux précédents. La fig. 12 montre deux exemples d'affectation de signe pour n pair, puis n impair, et la fig. 13 schématise le mineur d'un élément.

La valeur d'un déterminant est alors obtenue en choisissant une ligne ou une colonne quelconque, en faisant le produit de chaque élément de celle-ci par son cofacteur, et en sommant sur toute la ligne ou la colonne considérée. Le résultat est indifférent au choix de la ligne ou de la colonne. Le calcul d'un déterminant d'ordre n revient alors à celui de n déterminants d'ordre $(n-1)$, soit $n\ (n-1)$ déterminants d'ordre $(n-2)$, et donc à celui de

$$n\ (n-1)\ (n-2)\ ... \times 3 = \frac{1}{2} n!$$

déterminants d'ordre 2 que l'on sait aisément calculer. Le procédé est en général fastidieux dès que la valeur de n dépasse 6 ou 7.

Dans cette méthode, il faut choisir la ligne ou la colonne présentant le plus d'éléments nuls. En particulier, le déterminant d'une matrice triangulaire (*a fortiori* si elle est diagonale) est égal au produit des éléments de la diagonale.

Un certain nombre de propriétés permettent même de modifier l'écriture d'un déterminant sans en changer la valeur. Utilisées pour obtenir un grand nombre de zéros, ces propriétés sont la transcription de celles des formes n-linéaires alternées.

Propriété 1 : det A = det ${}^tA$.

Propriété 2 : permuter deux lignes — ou deux colonnes — revient à changer la valeur du déterminant en son opposée.

Propriété 3 : multiplier tous les éléments d'une même ligne — ou d'une même colonne — d'un déterminant par un nombre équivaut à multiplier la valeur du déterminant par ce nombre. Ce qui entraîne que det $(kA) = k^n$ det A, si n est l'ordre de A et $k \in K$ un scalaire.

Propriété 4 : ajouter à une ligne (resp. colonne) une combinaison linéaire des autres lignes (resp. colonnes) ne change pas la valeur d'un déterminant.

Les opérations entre déterminants se définissent alors simplement.

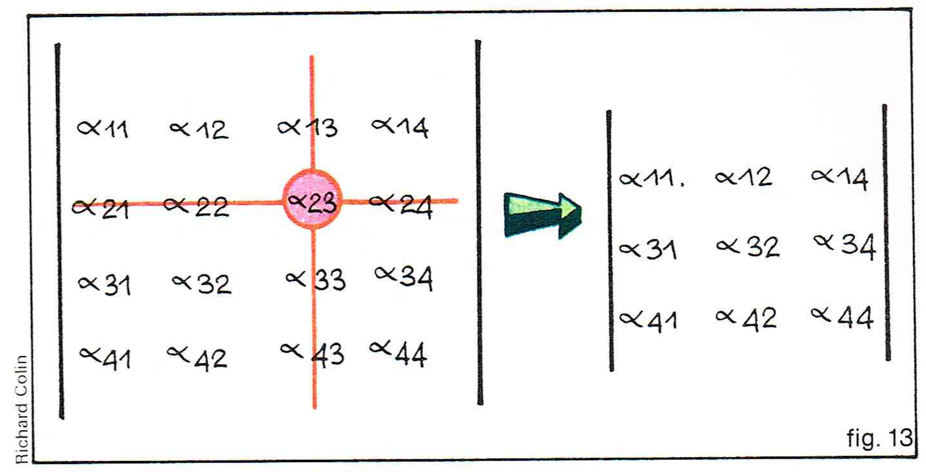
(1) La somme de deux déterminants D et D', ne différant que par une ligne (resp. colonne) de même position k, est un déterminant identique, mais dont la k-ième ligne (resp. colonne) est égale à la somme des k-ièmes lignes (resp. colonnes) de D et D'.

(2) Le produit de deux déterminants est le déterminant de la matrice produit.

De cette règle, on déduit une interprétation globale des déterminants. En effet, puisque

$$\det (AB) = (\det A) \times (\det B),$$

on a : det (AB) = det (BA). Or deux matrices semblables M et N sont liées par une relation $M = P^{-1} \cdot N{\cdot}P$, donc : $\det M = \det (P^{-1} \cdot N{\cdot}P) = \det (P^{-1} N) \cdot \det P = \det P^{-1} \cdot \det N \cdot \det P = \det P^{-1} \cdot \det P \cdot \det N$, soit $\det M = \det (P^{-1} P) \cdot \det N = \det N$ puisque $P^{-1} P = I_n$ et que $\det I_n = +1$.



▲ **Figure 13 :** ***caractérisation du mineur d'un élément.***

Il est donc plus clair, maintenant, de parler du déterminant d'un endomorphisme, et l'on peut énoncer comme suit.

Théorème. Soit $f \in \mathcal{L}$ (E) ; f est inversible si et seulement si la matrice de f, relativement à une base quelconque de E, a un déterminant non nul.

Applications. Les utilisations des déterminants sont nombreuses. Toutefois l'outil n'est pas toujours simple à manier. Ceci explique que les méthodes de calcul numérique utilisées dans les programmes de calcul scientifique soient basées sur des algorithmes plus proches des concepts linéaires. Toutefois deux applications pratiques méritent d'être mentionnées :

— la recherche du rang d'une matrice;

— l'orientation dans un espace vectoriel réel de dimension finie.

● Le théorème ci-dessous décrit totalement la première de ces questions.

Théorème. Si E et F sont deux espaces vectoriels de dimensions finies (non nécessairement égales) sur le corps K, si $f \in \mathcal{L}$ (E, F) et si A est la matrice de f relativement à deux bases choisies de E et F, alors le rang de cette application vérifie :

$$rg\ f = \dim f\ (E) = \dim \operatorname{Im} f = rg\ A = rg\ {}^tf = rg\ {}^tA$$

et ce nombre est encore le nombre maximal de vecteurs-colonnes — ou de vecteurs-lignes — linéairement indépendants que l'on peut extraire de A, soit aussi la dimension du plus grand mineur non nul qu'il est possible de trouver dans A.

● Dans le second problème, il s'agit d'une application très utile en physique, et qui consiste à distinguer deux classes de bases pour un espace vectoriel réel E, de dimension finie p, par l'intermédiaire d'une relation d'équivalence. Si $(e_i)_{i=1,\ ...,\ n}$ et $(f_j)_{j=1,\ ...,\ p}$ sont deux bases de E, on dira qu'elles sont équivalentes si le déterminant du système formé par l'une d'elles, par rapport à l'autre base, est positif; il s'agit bien d'une relation d'équivalence. Alors, pour une base B donnée, l'ensemble de toutes les bases de E peut être scindé en deux classes :

— les bases équivalentes à B, dites bases directes ou orientées positives;

— les autres bases, dites rétrogrades ou orientées négatives.

Mais les déterminants sont en réalité un outil d'algèbre multilinéaire dont les utilisations, fastidieuses dans le cas linéaire, peuvent souvent être évitées.

Les invariants

C'est encore une question de géométrie (la classification des courbes du second degré) qui est à l'origine du problème de la réduction des matrices, c'est-à-dire la recherche d'une base dans laquelle la matrice d'un endomorphisme prend une forme simple : diagonale, triangulaire ou encore forme de Jordan. Les études de A. Cayley, C. Hermite et J. J. Sylvester se sont révélées d'une rare fécondité par leur aspect unificateur des problèmes existants et leurs multiples applications; la théorie des représentations du groupe linéaire est en fait née de leurs travaux. C'est également la « théorie spectrale », et donc

le lien concret entre l'algèbre linéaire, l'analyse et l'algèbre tensorielle, qui est issue de leurs recherches.

Le premier problème est de déterminer s'il existe — et dans ce cas, de caractériser — des sous-espaces vectoriels de dimensions les plus petites possibles (le meilleur cas sera la dimension 1) qui soient invariants pour un endomorphisme f d'un espace vectoriel E de dimension finie p. Lorsqu'un tel sous-espace F est de dimension 1, il est engendré par un seul vecteur, non nul, dont l'image doit donc être dans F et par suite être un multiple de V. On est donc conduit à poser la définition suivante : un vecteur $V \in E$, non nul, est dit *vecteur propre* pour $f \in \mathcal{L}$ (E) s'il existe un scalaire k tel que $f(V) = kV$ **(1)**.

On dit que k est une valeur propre de f s'il existe un vecteur $V \in E$, non nul, vérifiant **(1)**.

A un vecteur propre V ne correspond qu'une seule valeur propre k, dite valeur propre associée à V ; tandis qu'à une valeur propre k correspond une infinité de vecteurs propres
[car, si $f(V) = kV$, alors $f(aV) = af(V) = akV = k(aV)$] dont l'ensemble, complété par le vecteur nul, forme un sous-espace vectoriel, que l'on appelle *sous-espace propre* associé à k. C'est un des sous-espaces invariants recherchés ; il est distinct de $\{0\}$, et sa dimension est donc au moins égale à 1. L'étude des dimensions des sous-espaces propres donne les résultats fondamentaux de la réduction des matrices.

Dans un espace du type $\mathbb{R}^p$, le but est d'abord la recherche des droites stables par f, c'est-à-dire dont les vecteurs aient des images homothétiques. Dans $\mathbb{R}^3$ par exemple, on cherche à « voir » si une transformation linéaire ne peut pas être ramenée à des dilatations le long de trois axes de coordonnées. La résolution analytique du problème tend souvent à faire oublier son aspect tant géométrique que linéaire. C'est pourquoi, à l'exception d'une méthode de calcul, on posera le plus souvent la question des sous-espaces propres, et non celle des valeurs propres.

Calcul des valeurs propres

Si k est valeur propre associée au vecteur propre V, on a : $f(V) = kV$, soit $f(V) - kV = 0$. Or $f(V) - kV$ est l'image du vecteur V par l'application linéaire $f - k \cdot \mathrm{Id}$, donc $V \in \mathrm{Ker}\,(f - k \cdot \mathrm{Id})$. Comme V est non nul, ce sous-espace doit être distinct de $\{0\}$, et donc l'application $f - k \cdot \mathrm{Id}$ ne doit pas être bijective et ne possède alors pas de réciproque. Par conséquent, si on se donne une base de E, et si par rapport à cette base A est la matrice de f, et I celle de l'application identité (cas de la dimension finie), les valeurs propres de f seront les solutions de l'équation :

$$\det(A - kI) = 0.$$

Si E est de dimension p, alors A est d'ordre p et $\det(A - kI)$ est un polynôme de degré p en k qu'on appelle *polynôme caractéristique* de l'endomorphisme f, ou de la matrice A.

Pour une valeur propre donnée k, l'équation matricielle $(A - kI)V = 0$ où V est un vecteur à p coordonnées permet de déterminer les sous-espaces propres. Puisque ce système homogène de p équations à p inconnues a une matrice non inversible, son rang r est strictement plus petit que p ; la différence entre p et r est la dimension du sous-espace propre. C'est en effet ce nombre qui caractérise le nombre d'inconnues non principales du système, donc son degré d'indétermination. On peut le voir encore d'une autre façon : $p = \dim \mathrm{Im}\,(f - kI) + \dim \mathrm{Ker}\,(f - kI)$ donc $\dim \mathrm{Ker}\,(f - kI) = p - r$.

Les propriétés algébriques des sous-espaces propres se résument dans le théorème qui suit.

Théorème. Si k_1 et k_2 sont deux valeurs propres distinctes d'une application $f \in \mathcal{L}$ (E), et si G (k_1) et G (k_2) sont les sous-espaces propres associés, on a :

$$G(k_1) \cap G(k_2) = \{0\}.$$

De plus si $k_1, k_2, ..., k_n$ sont toutes les valeurs propres distinctes de f, et si $V_1, V_2, ..., V_n$ sont des vecteurs propres qui leur sont respectivement associés, la famille de vecteurs $\{V_1, V_2, ..., V_n\}$ est libre.

Comme les systèmes libres de E ont au plus p éléments lorsque $\dim_k E = p$, il y a au plus p valeurs propres distinctes. Pour chercher s'il existe un certain nombre de sous-espaces de E, invariants par f, et dont la somme directe soit E, il faut et il suffit de chercher une base de E formée de vecteurs propres. Dans ce cas, l'endomorphisme f opère indépendamment dans chaque direction de l'espace à p dimensions, et chaque fois comme une homothétie.

La matrice A de f, relativement à cette base de vecteurs propres, prend une forme diagonale, car, pour tout vecteur V_i ($i = 1, 2, ..., p$) de la base, on a $f(V_i) = k_iV_i$ et le i-ième vecteur-colonne de A ne possède qu'une coordonnée non nulle, la i-ième, qui est égale à k_i, et A prend bien la forme ci-dessous.

$$\begin{pmatrix} k_1 & 0 & 0 & \dots & 0 \\ 0 & k_2 & 0 & \dots & 0 \\ 0 & 0 & k_3 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & k_n \end{pmatrix}$$

On parle dans ce cas de matrice (ou d'endomorphisme) *diagonalisable*. Une condition nécessaire et suffisante pour qu'il en soit ainsi est donnée par le théorème qui suit.

Théorème. Soit E de dimension finie sur le corps K. L'application $f \in \mathcal{L}$ (E) est diagonalisable si et seulement si toutes les racines du polynôme caractéristique sont dans K, et si, pour toute valeur propre, la dimension du sous-espace propre associé est égale à l'ordre de multiplicité de cette valeur propre dans le polynôme caractéristique.

En particulier, si f possède p valeurs propres distinctes toutes dans K, chacune ne peut être que racine simple du polynôme caractéristique. Cette condition n'est que suffisante ; elle n'est pas nécessaire.

Il n'est pas toujours possible de diagonaliser une matrice. La réduction à la forme triangulaire, moins intéressante, mais plus fréquemment possible, consiste à trouver une base de E telle que la matrice exprimée dans cette base soit triangulaire (inférieure ou supérieure). Ceci est réalisable dès que toutes les racines du polynôme caractéristique de f sont des éléments de K ; par exemple, si K = corps algébriquement clos *(cf. les Nombres)*, tout endomorphisme de E est « triangulaire ». La diagonale de la matrice triangulaire est alors formée par les valeurs propres.

La **forme réduite de Jordan**, forme de matrice triangulaire supérieure $A = (a_{ij})_{\substack{i=1,\dots,n \\ j=1,2,\dots,n}}$ telle que, parmi les éléments au-dessus de la diagonale, seuls les éléments $a_{i,\,i+1}$ puissent être non nuls (et dans ce cas égaux à 1) — les éléments diagonaux étant les valeurs propres, chacune comptée avec son ordre de multiplicité — peut être obtenue dès que l'endomorphisme est nilpotent (c'est-à-dire si une puissance de f est nulle, donc aussi les suivantes). L'exemple ci-dessous montre une forme réduite de Jordan (les éléments diagonaux ne sont pas nécessairement distincts).

$$\begin{pmatrix} k_1 & 1 & 0 & \dots & 0 & 0 \\ 0 & k_2 & 0 & \dots & 0 & 0 \\ 0 & 0 & k_3 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & 0 & \dots & k_{n-1} & 1 \\ 0 & 0 & 0 & \dots & 0 & k_n \end{pmatrix}$$

Un très célèbre résultat — le théorème de Cayley-Hamilton — est à l'origine de nombreux développements. Sous une forme particulièrement simple, il montre que si $f \in \mathcal{L}$ (E) admet $P(x) = a_0 + a_1x + ... + a_px^p$ pour polynôme caractéristique ($\dim_K E = p$), alors l'endomorphisme $a_0\mathrm{Id} + a_1f + ... + a_pf^p$ (où $f^k = f \circ f \circ ... \circ f$, k fois) est identiquement nul (tout élément de E a pour image le vecteur nul). Si A est la matrice de f relativement à une base quelconque, on a donc : $a_0I + a_1A + ... + a_pA^p = 0$.

On dit que le polynôme caractéristique est un polynôme annulateur de A ou de f.

La théorie des polynômes sur un anneau permet alors de dire (en dimension finie) que l'ensemble des polynômes annulateurs d'une matrice A (qui n'est pas vide puisqu'il contient le polynôme caractéristique) possède un élément de plus petit degré tel que tout autre élément en soit multiple : c'est le *polynôme minimal* de A (ou de f) dont le degré est au plus égal à p. Les racines de ce polynôme sont les mêmes que celles du polynôme caractéristique : les valeurs propres. Elles sont invariantes

— comme les polynômes d'ailleurs — par changement de base; leur ensemble est appelé *spectre de f*, noté Sp (f).

La notion de fonction de matrice, jusque-là limitée aux fonctions polynômiales, a pu alors être généralisée, et des expressions telles que le logarithme d'une matrice, une exponentielle de matrice, un sinus de matrices, etc. (dont l'usage s'avère primordial dans la résolution des systèmes d'équations différentiels), peuvent être considérées et calculées sans même avoir à utiliser de séries dans un espace vectoriel normé (en sortant donc du cadre linéaire sans nécessité expresse). La détermination du spectre d'une matrice apparaît donc comme fondamentale pour la connaissance d'un opérateur linéaire sur un espace vectoriel de dimension finie. Il suffit même souvent pour traiter certaines questions de localiser la plus grande des valeurs propres (par exemple, comparer son module à l'unité). Le théorème de Perron-Frobenius est une intéressante méthode de localisation; on peut en donner deux versions selon que la matrice est irréductible ou non.

On dit qu'une matrice carrée est *irréductible* si une même permutation effectuée sur les vecteurs-colonnes et sur les vecteurs-lignes permet de partitionner la matrice obtenue en une matrice de la forme :

$$\begin{pmatrix} M & O \\ P & Q \end{pmatrix} \quad \text{ou bien} \quad \begin{pmatrix} M & P \\ O & Q \end{pmatrix}$$

où M et Q sont deux sous-matrices carrées et O un bloc de zéros. Dans le cas contraire, la matrice est *réductible*. La figure 14a schématise la reconnaissance d'une matrice réductible tandis que la figure 14b fait apparaître un bloc de zéros diagonal qu'aucun échange simultané de lignes et colonnes ne peut placer dans l'une des positions en haut à droite ou en bas à gauche : la matrice est irréductible.

Il s'agit en fait de procéder à une permutation sur les vecteurs de la base choisie, c'est-à-dire dans le cas de la figure 14a, par exemple, de passer de la base (e_1, e_2, e_3, e_4) à la base (e_1, e_3, e_4, e_2). C'est un changement de base particulier caractérisé par une matrice de permutation qui n'est autre que la matrice identité où l'on a permuté les vecteurs-colonnes correspondant à ceux des vecteurs de base échangés; ainsi, dans le cas examiné, c'est la matrice :

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

et l'on vérifie par le calcul que :

$$\begin{pmatrix} 2 & 2 & 0 & 0 \\ 2 & 2 & 0 & 0 \\ 3 & 3 & 1 & 2 \\ 1 & 1 & 2 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & 0 & 2 & 0 \\ 3 & 1 & 3 & 2 \\ 2 & 0 & 2 & 0 \\ 1 & 2 & 1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

comme pour n'importe quel changement de base. L'utilisation de ces matrices est très courante dans l'étude des tableaux de données économiques pour étudier les interactions dans les systèmes d'échanges.

Théorème de Perron-Frobenius (cas irréductible)

Soit A une matrice carrée, à éléments positifs ou nuls, irréductible. Alors, il existe une valeur propre $k > 0$, plus grande que toute autre en valeur absolue, qui est simple (ordre de multiplicité égal à 1) et à laquelle correspond un vecteur propre dont toutes les composantes sont strictement positives; cette valeur propre est donnée par la formule :

$$k = \max_{\substack{v \geq 0 \\ v \neq 0}} \left[\min_i \frac{(Av)_i}{v_i} \right]$$

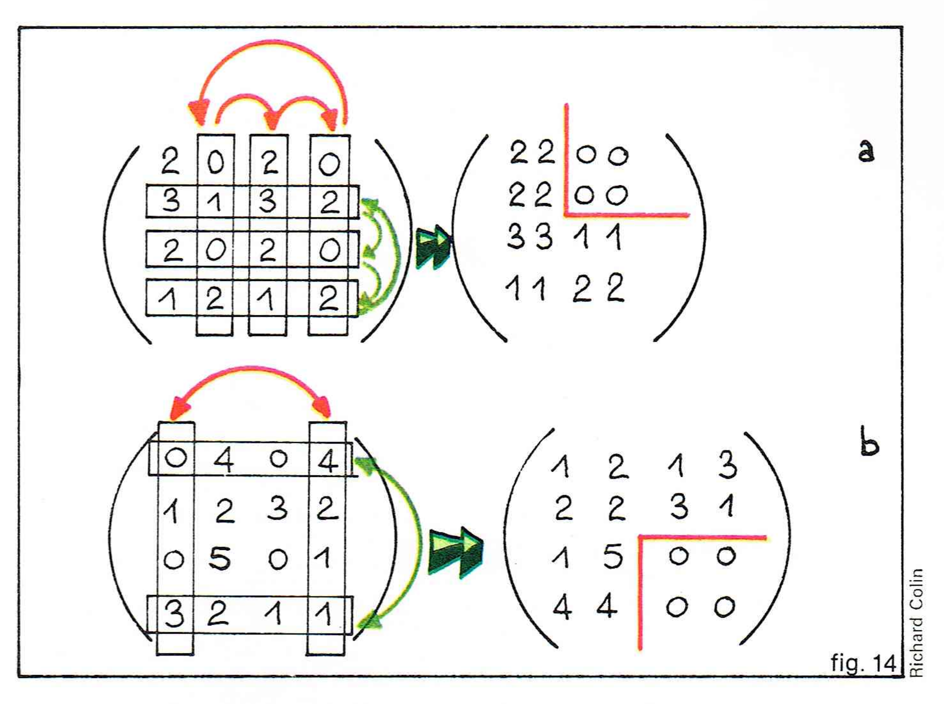
(l'indice i désigne la i-ième coordonnée du vecteur et $v \geqslant 0$ signifie qu'aucune composante de v n'est négative).

Dans le cas où A est réductible, la valeur propre de plus grand module peut éventuellement être nulle (dans ce cas, 0 est la seule valeur propre), multiple, et les vecteurs propres qui lui correspondent ont des coordonnées positives ou nulles.

Grâce à ce théorème, on a pu construire un algorithme qui permet de trouver, sans trop de calculs, une bonne valeur approchée de k même si l'ordre de la matrice A est grand.

Les matrices à éléments positifs dont on vient de parler se rencontrent fréquemment : ainsi en probabilité où l'on utilise les *matrices stochastiques* qui sont telles que la somme des éléments (non négatifs) d'une même ligne soit égale à 1.

La plus grande valeur propre est alors $k = 1$ et le vecteur $(1, 1, 1, ..., 1)$ est un vecteur propre associé. Cette propriété est équivalente à la définition.



▲ **Figure 14 — a, matrice réductible :** ***une permutation des lignes et des colonnes permet d'isoler un bloc carré de zéros, non situé sur la diagonale;*** **b, matrice irréductible :** ***le bloc de zéros ne peut être que diagonal.***

Formes quadratiques

Si $f : E \times E \to K$ est une forme bilinéaire sur E, on dit que $Q(x) = f(x, x)$ est la *forme quadratique associée à f*. La dénomination est claire puisque

$$Q(ax) = f(ax, ax) = a^2 f(x, x) = a^2 Q(x).$$

On dit que Q est *positive* si $Q(x, x) \geqslant 0$ pour tout $x \in E$; elle est *définie positive* si de plus $Q(x, x) = 0$ n'a lieu que pour $x = 0$. A toute forme bilinéaire f, sur E, on associe la matrice $M = (m_{ij})_{\substack{i=1,2,...,p \\ j=1,2,...,p}}$ relativement à une base $(e_i)_{i=1,...,p}$ par : $m_{ij} = f(e_i, e_j)$. On définit ainsi un isomorphisme, mais non canonique, entre l'ensemble des formes bilinéaires sur E et l'ensemble $\mathcal{M}_p(K)$ des matrices carrées d'ordre p.

Si X est le vecteur-colonne représentant x et Y celui représentant y dans la base $(e_i)_{i=1,...,p}$, on a alors :

$$f(x, y) = {}^tY \cdot {}^tM \cdot X = {}^tX \cdot {}^tM \cdot Y$$

Lorsque f est symétrique, soit $f(x, y) = f(y, x)$, on a :

$$Q(x + y) = Q(x) + Q(y) + 2 f(x, y);$$

donc à toute forme quadratique Q on peut associer une forme bilinéaire symétrique f, dite *forme polaire* de Q par :

$$f(x, y) = \frac{1}{2} [Q(x + y) - Q(x) - Q(y)]$$

L'application $f \to Q$ est donc une bijection de l'ensemble des formes bilinéaires symétriques dans l'ensemble des formes quadratiques sur E. On peut donc écrire :

$$Q(x) = {}^tX \cdot M \cdot X$$

puisque, si M est la matrice de f, on peut dire que M est la matrice de Q; cette matrice est symétrique, car :

$$m_{ij} = f(e_i, e_j) = f(e_j, e_i) = m_{ji}, \text{ donc } M = {}^tM.$$

Le problème de la réduction des matrices symétriques à la forme diagonale est donc le même que celui de la transformation — par changement de base — d'une forme quadratique en une somme algébrique de carrées. La représentation globale des courbes du second degré a trouvé là son terme après les énumérations fastidieuses longtemps enseignées. Une étude simple de la question à l'aide de la notion de dualité conduit aux deux résultats suivants.

Loi d'inertie de Sylvester

Soit f une forme bilinéaire symétrique réelle sur un espace vectoriel réel E de dimension p, et Q la forme quadratique associée. Alors, il existe une base $(v_i)_{i=1,\ldots,p}$ telle que $f(v_i, v_j) = \delta_{ij}$ (symbole de Kronecker) — base *orthonormale* relativement à f — et un couple unique $(m, n) \in \mathbb{N} \times \mathbb{N}$ tels que, dans cette base, on ait :

$$Q(x) = \sum_{i=1}^{i=m} x_i^2 - \sum_{j=m+1}^{j=m+n} x_j^2 \quad \text{avec } m + n = \text{rg}(f).$$

Théorème. Pour toute matrice symétrique réelle M d'ordre p, il existe une matrice T réelle d'ordre p telle que $T^{-1} = {}^tT$ (on dit que T est *orthogonale*) et que la matrice $T^{-1} \cdot M \cdot T$ soit diagonale.

Ainsi la forme bilinéaire symétrique définie sur $\mathbb{R}^3$:

$$f(x, y) = x_1y_1 + x_2y_2 + x_3y_3 - 2(x_2y_3 + x_3y_2 + x_3y_1 + x_1y_3 + x_1y_2 + x_2y_1)$$

est représentée relativement à la base canonique par la matrice

$$M = \begin{pmatrix} 1 & -2 & -2 \\ -2 & 1 & -2 \\ -2 & -2 & 1 \end{pmatrix}, \quad \text{symétrique.}$$

La forme quadratique associée :

$$Q(x) = x_1^2 + x_2^2 + x_3^2 - 4(x_2x_3 + x_3x_1 + x_1x_2)$$

peut se réduire à la forme :

$$Q(x) = [x_1 - 2(x_2 + x_3)]^2 + \frac{3}{2}(x_2 - x_3)^2 - \frac{9}{2}(x_2 + x_3)^2$$

tandis que la matrice M se diagonalise sous la forme :

$$\begin{pmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & -3 \end{pmatrix}$$

L'exemple de la matrice M montre la conservation de la somme des éléments diagonaux, après changement de base ; on peut aussi le constater sur l'exemple de matrice réductible vu plus haut. Il s'agit encore d'un invariant, appelé *trace* d'un endomorphisme f, noté Tr (f), égal à la somme de ses valeurs propres (l'espace E étant de dimension finie, il s'agit là d'une somme finie), ou encore à la somme des éléments diagonaux de la matrice représentant f dans une base quelconque.

Algèbre linéaire et groupes de transformation

Dans ce paragraphe, on va donner une classification élémentaire, par la structure de groupe, de quelques types de transformations linéaires d'un espace vectoriel E de dimension finie sur un corps K.

Le *groupe linéaire* $GL_K(E)$, ensemble des endomorphismes injectifs (ou automorphismes) de E, muni de la loi de composition des applications, est le plus large de ceux-ci. C'est à l'intérieur de ce groupe que l'on caractérise deux sous-groupes particuliers : le *groupe orthogonal* et le *groupe unitaire*.

Adjoint d'un endomorphisme

Sur E, soit S une forme bilinéaire symétrique telle que les deux applications linéaires partielles soient injectives (on dit que S est *non dégénérée*). Alors, pour toute application $f \in \mathcal{L}(E)$, il existe un endomorphisme unique f^* tel que :

$$S[f(x), y] = S[x, f^*(y)] \text{ pour tout couple } (x, y) \in E \times E.$$

Cet endomorphisme s'appelle l'*adjoint* de f. Le lien de cette notion avec la dualité, que la notation sous-entend, est plus net lorsqu'on sait que, relativement à toute base $(e_i)_{i=1,\ldots,n}$ telle que $S(e_i, e_j) = \delta_{ij}$ (symbole de Kronecker) — ou base orthonormale relativement à S — la matrice M^* de f^* et celle, M, de f sont liées par :

$$M^* = {}^tM \quad \text{si} \quad K = \mathbb{R} \quad \text{et} \quad M^* = {}^t\overline{M} \quad \text{si} \quad K = \mathbb{C}$$

($\overline{M}$ désigne la matrice dont les éléments sont les conjugués de ceux de M). Les principales propriétés de l'adjoint sont résumées dans le tableau I.

Tableau I

Propriétés de l'adjoint d'un endomorphisme

Corps de base K = R		Corps de base K = C
$(f + g)^*$	=	$f^* + g^*$
$(f \circ g)^*$	=	$g^* \circ f^*$
$(f^*)^*$	=	f
$\text{rg}(f^*)$	=	$\text{rg}(f)$
$(\alpha f)^* = \alpha f^*$		$(\alpha f)^* = \overline{\alpha} f^*$
$\text{Tr}(f^*) = \text{Tr}(f)$		$\text{Tr}(f^*) = \overline{\text{Tr}(f)}$
$\det(f^*) = \det(f)$		$\det(f^*) = \overline{\det(f)}$

▶ **Tableau I :** *propriétés de l'adjoint d'un endomorphisme.*

Transformation orthogonale

On suppose que $K = \mathbb{R}$, et on cherche à caractériser les transformations linéaires qui conservent les distances. On dit que $f \in \mathcal{L}(E)$ est *orthogonale relativement à S* si, pour tout couple $(x, y) \in E \times E$, on a :

$$S[f(x), f(y)] = S(x, y).$$

Si Q est la forme quadratique associée à S, ceci revient à $Q[f(x)] = Q(x)$, pour tout $x \in E$. Alors f et son adjointe vérifient $f \circ f^* = f^* \circ f = \mathrm{Id}$.

Par suite f est inversible et $f^{-1} = f^*$. Ceci justifie donc la définition suivante.

Une matrice A est *orthogonale* si elle vérifie :

$$A \times {}^tA = {}^tA \times A = I \quad \text{ou encore : } A^{-1} = {}^tA.$$

Leurs propriétés sont nombreuses.

Propriété 1. Si A est orthogonale, alors $|\det A| = +1$. Lorsque $\det A = +1$, on dit que la matrice est *droite* ; si $\det A = -1$, la matrice est dite *gauche*.

Propriété 2. Si A est orthogonale, toute valeur propre k de A vérifie $|k| = 1$.

En particulier, les valeurs propres d'une matrice orthogonale droite d'ordre 3 sont : 1, e^{ix} et e^{-ix} pour une certaine valeur réelle de x ; c'est-à-dire que l'application linéaire associée est une rotation d'angle x autour d'un axe Δ, dont tous les points sont invariants (ce qui explique la valeur propre 1). L'application linéaire associée à une matrice orthogonale gauche d'ordre 3 est une symétrie par rapport soit à un plan, soit à un point.

Les transformations orthogonales relativement à S décrivent un sous-groupe de $GL_{\mathbb{R}}(E)$ qu'on note $O(n, \mathbb{R})$ lorsque $S(x, y) = \sum_{i=1}^{i=p} x_i y_i$ (produit scalaire). Lorsque E est un espace vectoriel normé par le produit scalaire usuel — on dit que E est un *espace euclidien* — les transformations orthogonales de E sont les isométries de E.

Forme hermitienne

Le but est ici de reprendre les notions qui ont permis la construction du groupe orthogonal, mais lorsque le corps de base est $\mathbb{C}$.

Si E est un espace vectoriel sur $\mathbb{C}$, une application $h : E \times E \to \mathbb{C}$ est dite *forme hermitienne* si les relations

$$\begin{cases} h(ax + x', y) = ah(x, y) + h(x', y) \\ h(a, by + y') = \bar{b}h(x, y) + h(x, y') \\ h(x, y) = \overline{h(y, x)} \end{cases}$$

sont vérifiées pour tous éléments $x \in E$, $y \in E$, $x' \in E$, $y' \in E$, $a \in \mathbb{C}$ et $b \in \mathbb{C}$.

Cette notion étend celle de forme bilinéaire symétrique réelle au cas complexe. Dans ce cas, $h(x, x)$ est réel et l'application $x \to h(x, x)$ s'appelle forme quadratique hermitienne associée à h.

Soit une base $(e_i)_{i=1,\dots,p}$ de E ; la matrice $H = (h_{ij})$ définie par :

$h_{ij} = h(e_i, e_j)$ caractérise la forme hermitienne h, puisque, si $x = \sum_{i=1}^{i=p} x_i e_i$ et $y = \sum_{j=1}^{j=p} y_j e_j$, on a $h(x, y) = \sum_{i=1}^{i=p}\sum_{j=1}^{j=p} x_i \bar{y}_j h_{ij}$.

Cette matrice vérifie ${}^tH = \bar{H}$ puisque

$$h_{ji} = h(e_j, e_i) = \overline{h(e_j, e_i)} = \overline{h_{ji}}.$$

L'inégalité de Schwarz : $|h(x, y)|^2 \leqslant h(x, x) \times h(y, y)$ est la propriété de base des formes hermitiennes.

Une forme hermitienne est dite *non dégénérée* si la matrice H associée est inversible, soit si $\det H \neq 0$ (lorsque $h(x, x) \geqslant 0$ pour tout $x \in E$, on dit que la forme est positive). Alors, la forme quadratique Q associée à h définit une application $N : E \to \mathbb{R}^+$, $N : x \to \sqrt{Q(x)}$ qui vérifie :

$$N(x) = 0 \Leftrightarrow x = 0$$
$$N(\alpha x) = |\alpha|\, N(x) \quad \text{si } \alpha \in \mathbb{C} \text{ et } x \in E$$
$$N(x + y) \leqslant N(x) + N(y) \quad \text{si } x \in E \text{ et } y \in E$$

et qui s'appelle une *norme* sur E. Un espace vectoriel complexe E, de dimension finie, muni de la norme issue d'une forme hermitienne non dégénérée positive est appelé *espace hermitien*. Alors, à toute application $f \in \mathcal{L}(E)$ correspond un endomorphisme unique f^* (adjoint de f) tel que : $h[f(x), y] = h[x, f^*(y)]$ pour tout couple $(x, y) \in E \times E$, dont les propriétés sont résumées dans le tableau I.

Les matrices M de f et M* de f^* relativement à toute base orthonormale de E sont adjointes : $M^* = {}^t\bar{M}$.

Groupe unitaire

Comme dans le cas du groupe orthogonal, on peut montrer que, sur un espace hermitien E, de dimension finie, tout endomorphisme bijectif $u \in \mathcal{L}(E)$ tel que $u = u^*$ vérifie aussi : $h[u(x), u(y)] = h(x, y)$ et $N[u(x)] = N(x)$ si h est la forme hermitienne qui définit la norme N sur E.

Un tel endomorphisme s'appelle *automorphisme unitaire* ; cette dénomination provient du fait que le nombre complexe $\det u$ est de module égal à 1. Une matrice carrée d'ordre n, A, à coefficients complexes, telle que

$$A^* \cdot A = A \cdot A^* = I, \quad \text{soit } A^* = {}^t\bar{A} = A^{-1},$$

est donc appelée *matrice unitaire*. En particulier, une matrice unitaire réelle est une matrice orthogonale.

Les automorphismes unitaires d'un espace hermitien de dimension finie forment un sous-groupe de $GL_{\mathbb{C}}(E)$, le *groupe unitaire*, que l'on désigne par $U(n, \mathbb{C})$.

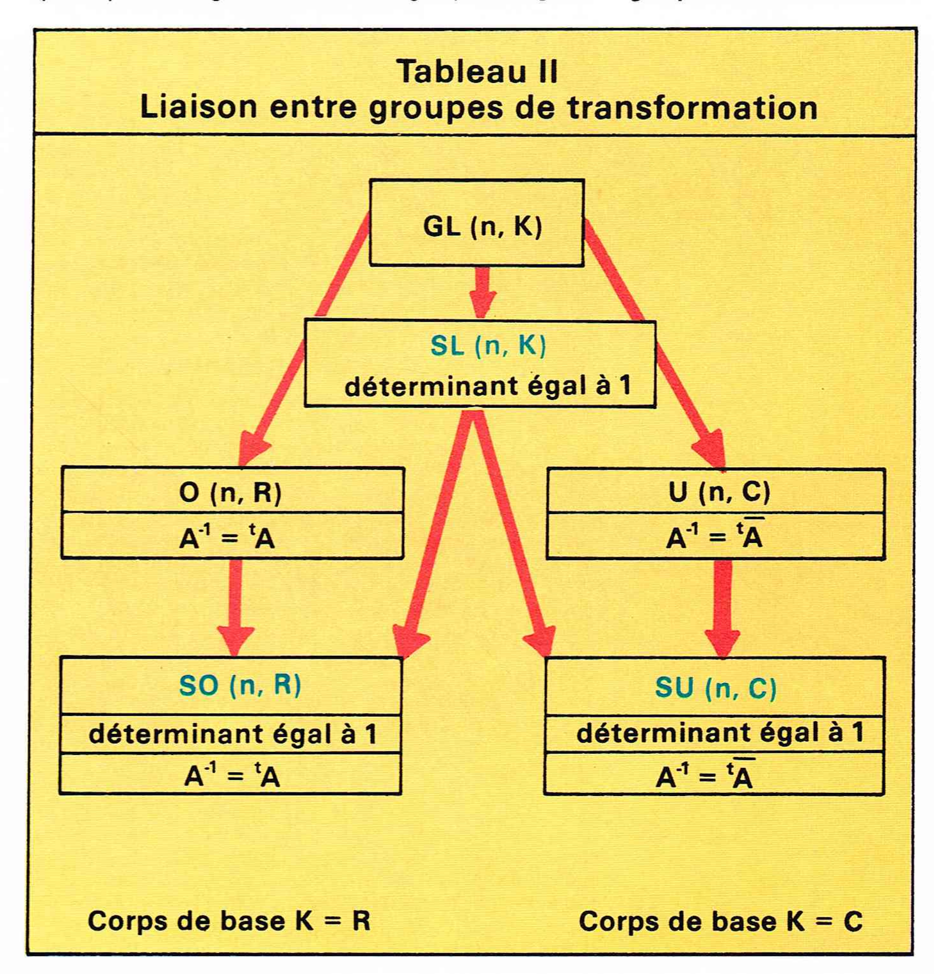
Parmi les opérateurs unitaires, ceux dont le déterminant est exactement $+1$ forment un sous-groupe de $U(n, \mathbb{C})$, le *groupe spécial unitaire* noté $SU(n, \mathbb{C})$; de même que parmi les transformations orthogonales, celles dont le déterminant vaut $+1$ forment un sous-groupe de $O(n, \mathbb{R})$, le *groupe spécial orthogonal* noté $SO(n, \mathbb{R})$. Plus généralement, dans le groupe $GL_K(E)$ les endomorphismes dont le déterminant est égal à $+1$ (qui sont donc des automorphismes de E) forment un sous-groupe, le *groupe spécial linéaire* $SL(n, K)$.

Le tableau II schématise les liens entre ces groupes d'opérateurs (les flèches rouges symbolisent une inclusion).

L'extension pour $K = \mathbb{C}$ des résultats obtenus pour les endomorphismes symétriques est le théorème : pour toute matrice hermitienne H, il existe une matrice unitaire U telle que la matrice $U^{-1} \cdot H \cdot U$ soit diagonale.

On dit aussi que toute matrice hermitienne est diagonalisable dans le groupe unitaire, de même que toute matrice symétrique est diagonalisable dans le groupe orthogonal.

▼ **Tableau II :** ***liaisons entre groupes de transformation.***



Espaces vectoriels de dimension infinie

La structure d'espace vectoriel est très riche, les propriétés des opérateurs linéaires (plus spécialement ceux des groupes de transformation cités ci-dessus) offrent des applications extrêmement diverses, que ce soit pour d'autres développements mathématiques ou pour des applications en physique, en statistiques, en économie, etc. La simplicité, mais en même temps la richesse du concept de linéarité, lui donnent cette importance fondamentale, et l'utilisation du calcul matriciel, par exemple, n'en est certes pas le plus petit aspect. Il est tout de même des barrières qu'il faut poser dès que l'on cherche à décrire et étudier des espaces vectoriels qui ne sont plus de dimension finie. L'outil matriciel, ainsi, n'est plus disponible; mais ce n'est pas la seule limitation. La première distinction est celle qui intervient sur les bases et plus encore sur la dimension d'un espace vectoriel, car bien évidemment la description globale d'un espace vectoriel par un nombre fini de ses éléments est exceptionnellement féconde. Tout espace vectoriel de dimension p sur un corps K est isomorphe à K^p, c'est-à-dire qu'il n'y a en fait qu'une seule structure d'espace à p dimensions sur K; on ne peut rien dire d'analogue en dimension infinie. Puis, dès que l'on s'intéresse aux applications linéaires, les différences sont encore plus sensibles; une application linéaire n'est plus bijective dès qu'elle est seulement injective ou surjective. Les applications développées sur le concept de rang ne sont plus utilisables lorsque le sous-espace vectoriel f (E) est de dimension infinie. Le fossé est très large pour les questions de dualité, et tout le schéma de « réciprocité », basé sur l'application de E dans E^{**}, ne peut plus être utilisé. En particulier, il n'est plus vrai que ${}^t({}^tf) = f$, ou que rg $({}^tf) =$ rg f.

Bien entendu, tous les problèmes des invariants doivent être pensés différemment quoique les concepts de base, valeurs propres et vecteurs propres restent les mêmes. La pratique de leur détermination ne peut plus être systématisée par les déterminants et les systèmes d'équations linéaires; ce sont des considérations spécifiques à chaque problème particulier qui permettent de le résoudre. Plus généralement, c'est toute la théorie spectrale qui est à reconstruire, mais il faut cette fois de nouveaux outils.

C'est avec la topologie et l'analyse moderne que l'on a pu poursuivre l'étude des espaces vectoriels de dimension infinie : espaces vectoriels topologiques et surtout espaces de Hilbert. Les espaces de fonctions ont, en effet, fourni les plus nombreux exemples d'ensembles sur lesquels on définit aussi bien une structure linéaire qu'une structure métrique avec un produit scalaire et même une norme : ensemble des fonctions continues sur un intervalle fermé borné de $\mathbb{R}$; ensemble des fonctions intégrables au sens de Lebesgue ou bien surtout ensemble des fonctions de carré intégrable, au sens de Lebesgue, noté $\mathfrak{L}^2$, qui par passage à l'ensemble quotient par une relation d'équivalence donne l'espace de Hilbert L^2.

Nouvelles tendances de l'algèbre linéaire

La notion d'espace vectoriel sur un corps cède la place à celle, plus générale, de *module* sur un anneau A ou encore A-module. La définition s'obtient en remplaçant, dans celle d'un espace vectoriel sur un corps K, le corps K par l'anneau A. On voit ainsi qu'un anneau unitaire peut être envisagé comme un module sur lui-même.

On définit de même les bases d'un module mais, et là se trouve une première différence fondamentale, tout module ne peut être muni d'une base; par exemple le $\mathbb{Z}$ module $\mathbb{Z}/2\mathbb{Z}$. Un module qui admet une base est dit *module libre*.

Module libre

Les morphismes de modules sont encore appelés applications linéaires. Les différentes propriétés des anneaux entraînent une classification variée des modules : selon que l'anneau est intègre, principal, etc.

Un module M sur un anneau intègre A est dit *sans torsion* si $ax = 0$ ($a \in A$ et $x \in M$) implique $a = 0$ ou $x = 0$. Cette classe de modules s'est révélée avoir de nombreuses propriétés. Ainsi, si A est un anneau principal (anneau intègre, unitaire et dont tout idéal est principal, c'est-à-dire engendré par les multiples d'un seul élément), et si M est un A-module de type fini (qui admet un système générateur fini) et sans torsion, alors M est un module libre. La seconde hypothèse est ici fondamentale, car, par exemple, $\mathbb{Q}$ est un $\mathbb{Z}$ module sans torsion, mais non libre.

Soit A un anneau intègre et K le corps obtenu par les quotients d'éléments de A (corps des fractions de A). Si l'on a un A-module libre M, on peut considérer l'espace vectoriel sur K, engendré par M. Alors pour tout sous-module M' de M, on peut chercher le nombre maximal d'éléments linéairement indépendants dans M'; ce nombre — dimension du sous-espace engendré par M' — est appelé le *rang* de M'. Dans le cas d'un sous-module libre dont une base possède p éléments, le rang est alors égal à p. Le résultat fondamental s'énonce par le théorème qui suit.

Théorème. Soit A un anneau principal, M un A-module libre de rang fini p. Si M' est un sous-module de M, M' est libre et de rang fini $r \leqslant p$. D'autre part, on peut trouver une base $(e_1, e_2, ..., e_p)$ de M, un entier $r \leqslant p$ et des éléments non nuls $a_1, a_2, ..., a_r$ de A tels que :

$(a_1e_1, a_2e_2, ..., a_re_r)$ soit une base de M'
a_i divise a_{i+1} pour $i = 1, 2, ..., r-1$.

L'importance de cette structure de module est très grande. Les anneaux nœthériens généralisant les anneaux de Dedekind, dans de multiples applications de l'algèbre (ainsi en géométrie algébrique), jouent un rôle fondamental. Il y a là, comme pour les espaces vectoriels, un effort pour la linéarisation des problèmes. Pour ne pas rentrer en détail dans ce domaine qui demanderait de longs développements explicatifs, on définira ici seulement les modules et anneaux nœthériens.

Soit A un anneau; un A-module M est dit *nœthérien* si tout sous-module de M est de type fini. Toute suite croissante de sous-modules est alors stationnaire, c'est-à-dire reste fixe au-delà d'un certain rang. Par conséquent, toute famille non vide de sous-modules de M possède un élément maximal. On dit qu'un anneau A est nœthérien si, lorsqu'on le considère comme un A-module, c'est un anneau nœthérien; un anneau principal est donc nœthérien. Il s'agit là d'un outil très important pour l'algèbre non commutative et la théorie des nombres. De nombreux développements sont dus à des mathématiciens contemporains : Eilenberg, Krull, Mac-Lane.

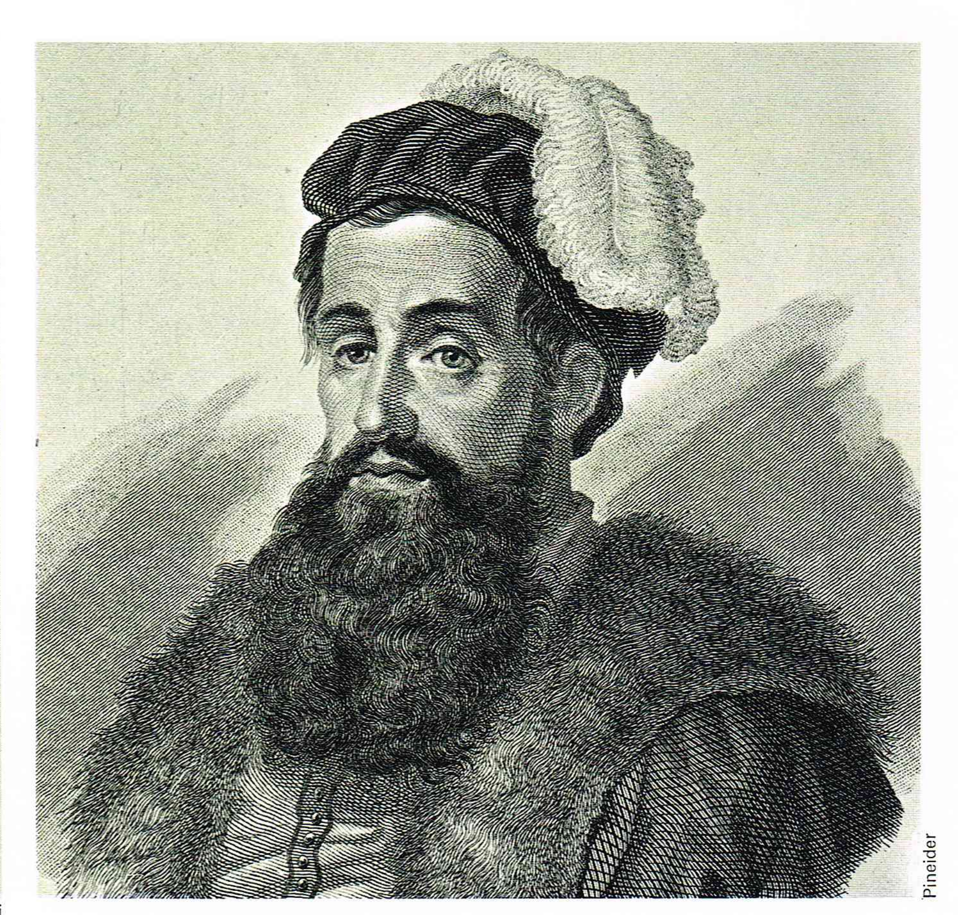
Il faut citer encore l'*algèbre homologique* dont l'origine vient des théories de l'homologie et de l'homotopie, et dont une importante application est justement de déterminer la limite des propriétés des modules par rapport à celles des espaces vectoriels; pour cela, on définit des modules associés à un A-module donné M, et qui se réduisent à $\{0\}$ lorsque A est un corps. Les travaux de Henri Cartan et d'Alexandre Grothendieck ainsi que de Samuel Eilenberg ont largement fait progresser cette branche que l'on peut maintenant considérer comme autonome.

BIBLIOGRAPHIE

ARTIN E., *Algèbre géométrique,* Gauthier-Villars. - BOURBAKI N., *Algèbre,* I à III, VI, VII, IX, Hermann; *Éléments d'histoire des mathématiques,* Hermann. - DIEUDONNÉ J., *Algèbre linéaire et Géométrie élémentaire,* Hermann. - FÉLIX L., *Exposé moderne des mathématiques élémentaires,* Dunod. - GODEMENT R., *Cours d'algèbre,* Hermann. - KEMENY, SNELL, THOMSON, *Algèbre moderne et Activités humaines,* Dunod. - QUEYSANNE M., *Cours d'algèbre,* Armand Colin.

LES ÉQUATIONS ALGÉBRIQUES

La résolution des équations algébriques a certainement été l'origine de développements parmi les plus importants des mathématiques. Les méthodes de recherche des solutions de telles équations, dès que le degré est supérieur à l'unité, ont fait intervenir de tout autres concepts que ceux utilisés pour les équations linéaires. Longtemps, les algébristes ont cherché à inclure l'extraction des racines n-ièmes parmi les opérations élémentaires de l'algèbre. En cherchant dans cette voie, ils parvinrent à résoudre les équations de degrés 2, 3 et 4, mais on dut attendre le XIX[e] siècle pour dresser une résolution d'ensemble des équations par la notion des groupes de Galois dont une très importante application est le théorème d'Abel sur la résolubilité des équations par radicaux.



Pedicini

Pineider



La méthode affine

Les équations du premier degré, grâce à l'outil formel de l'algèbre linéaire, se résolvent sans aucun problème. Sur un ensemble E, muni de deux lois — notées additivement et multiplicativement — qui lui confèrent une structure d'anneau, la recherche des éléments X tels que AX = B, où A et B sont donnés dans E, revient à la recherche de l'élément A^{-1}, inverse de A, lorsqu'il existe, ou sinon à un problème déjà expliqué dans le cadre des espaces vectoriels et qui se transpose aisément dans le cadre des modules sur un anneau : on interprète l'élément A comme une application linéaire de E dans lui-même, et l'on recherche tout d'abord l'ensemble des éléments images puis leurs « translatés » par l'élément B.

La méthode affine, qui est ainsi brièvement résumée, couvre entièrement la résolution des problèmes linéaires. Les techniques matricielles ont permis de dresser des méthodes algorithmiques complètes de résolution des équations du premier degré.

Le second degré

Le saut est alors très grand pour passer aux équations de degré supérieur; jusqu'ici, les équations linéaires n'utilisent que les opérations fondamentales définissant la structure, tandis que, dès le second degré, on voit apparaître l'extraction de radicaux. Dans l'équation

$$az^2 + bz + c = 0 \quad \text{où } a \neq 0,$$

l'étude du discriminant $\Delta = b^2 - 4\,ac$ permet de distinguer les trois cas bien connus (a, b, c sont des nombres réels).

Lorsque $\Delta > 0$, on peut écrire :

$$a\left(z - \frac{-b+\sqrt{\Delta}}{2\,a}\right)\left(z - \frac{-b-\sqrt{\Delta}}{2\,a}\right) = 0$$

d'où l'on déduit l'existence de deux racines réelles distinctes.

Lorsque $\Delta = 0$, on a alors $a\left(z + \dfrac{b}{2\,a}\right)^2 = 0$, et l'équation possède une racine double réelle.

Lorsque $\Delta < 0$, on a recours à la construction des nombres complexes pour écrire :

$$a\left(z - \frac{-b+i\sqrt{\Delta}}{2\,a}\right)\left(z - \frac{-b-i\sqrt{-\Delta}}{2\,a}\right) = 0.$$

C'est à l'Italien R. Bombelli que l'on doit le premier exposé sur les nombres complexes, contenant des règles de calcul; cet exposé contient d'ailleurs en germe de nombreux concepts d'algèbre linéaire.

Les deux racines sont alors deux nombres complexes conjugués. Les règles de calcul sur le corps $\mathbb{C}$ permettent la généralisation immédiate du procédé de résolution de $az^2 + b + c = 0$ lorsque a, b, c sont des nombres imaginaires.

Équations de degrés 3 et 4

La forme générale : $ax^3 + bx^2 + cx + d = 0$ **(1)** de l'équation du troisième degré peut se ramener à celle, plus simple, **(2)** $z^3 + pz + q = 0$ où p et q sont des expressions rationnelles de a, b, c et d :

$$p = \frac{-b^2 + 3\,ac}{3\,a^2} \qquad \text{et} \qquad q = \frac{2\,b^3 - 9\,abc + 27\,a^2d}{27\,a^3}$$

Il suffit pour cela de poser $z = x + \dfrac{b}{3\,a}$ après avoir divisé les deux membres de l'équation de départ par le terme a, supposé non nul (sinon l'équation serait de degré 2).

C'est au début du XVI[e] siècle que Scipion del Ferro trouve alors la formule de résolution :

$$z = \sqrt[3]{-\frac{q}{2} + \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}} + \sqrt[3]{-\frac{q}{2} - \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}}$$

Cette sensationnelle découverte relança, à travers des querelles d'écoles, les recherches sur les équations. Hyeronimus Cardan (1501-1576), après Nicolas Tartaglia, en 1545, publie les formules de résolution qu'il complète en formulant quelques observations. Il remarque ainsi que les équations du troisième degré peuvent avoir trois racines; que la somme des racines de **(1)** est toujours égale à $-\dfrac{b}{a}$. La méthode de Cardan pour résoudre **(2)** consiste à poser $z = y + t$, ce qui donne

$$y^3 + t^3 + (3\,yt + p)\,(y + t) + q = 0.$$

▲ A gauche, une page manuscrite de l'*Algèbre *de R. Bombelli, mathématicien italien à qui l'on doit le premier exposé sur les nombres complexes, contenant des règles de calcul.
A droite,* Nicolas Tartaglia *à qui l'on doit notamment les formules de résolution pour l'équation générale du troisième degré.



▲ *Frontispice de l'*Algèbre nouvelle *de F. Viète, maître des Requêtes de l'Hôtel du Roi (1540-1603) ; il fut un des grands mathématiciens de son siècle.*

Collection Viollet

On cherche alors les racines de cette nouvelle équation telles que $3\,yt + p = 0$; on a donc le système :

$$\begin{cases} y^3 + t^3 = -q \\ yt = -\dfrac{p}{3} \end{cases}$$

En posant alors $\alpha = y^3$ et $\beta = t^3$, on a :

$$\alpha + \beta = -q \quad \text{et} \quad \alpha\beta = -\frac{p^3}{27}$$

Les nombres α et β sont donc racines de l'équation du second degré en u :

(3) $$u^2 + qu - \frac{p^3}{27} = 0.$$

Par suite si γ représente l'une des trois déterminations du nombre $\sqrt[3]{\alpha}$, et $1, j, j^2$ les trois racines cubiques de l'unité, les trois valeurs possibles pour y sont γ, $j\gamma$ et $j^2\gamma$. Puis de $t = -\dfrac{p}{3\,y}$, on obtient les trois valeurs possibles correspondantes pour t :

$$-\frac{p}{3\,\gamma}, \quad -\frac{pj^2}{3\,\gamma} \quad \text{et} \quad -\frac{pj}{3\,\gamma}.$$

Par conséquent, les solutions de **(2)** sont :

$$z_1 = \gamma - \frac{p}{3\,\gamma}, \quad z_2 = j\gamma - \frac{pj^2}{3\,\gamma}, \quad z_3 = j^2\gamma - \frac{pj}{3\,\gamma}.$$

La discussion porte sur les racines de l'équation **(3)**, dont le discriminant est : $\Delta = 4\left(\dfrac{q^2}{4} + \dfrac{p^3}{27}\right)$. Lorsque $\dfrac{q^2}{4} + \dfrac{p^3}{27} > 0$ pour l'équation **(2)**, on a une racine réelle et deux autres complexes conjuguées :

— lorsque $\dfrac{q^2}{4} + \dfrac{p^3}{27} < 0$, l'équation **(2)** admet trois racines réelles ;

— lorsque $\dfrac{q^2}{4} + \dfrac{p^3}{27} = 0$, on trouve deux racines réelles, dont une double.

Malgré son intérêt, cette méthode est peu employée ; on lui préfère des méthodes numériques, faisant appel, par exemple, à la trigonométrie.

C'est un élève de Cardan, Luigi Ferrari, qui, en 1545, énonce la règle pour résoudre l'équation générale du quatrième degré :

$$x^4 + px^3 + qx^2 + rx + s = 0 \qquad \textbf{(4)}$$

Auparavant, l'équation bicarrée $y^4 + my^2 + n = 0$ avait été longuement étudiée et Cardan avait remarqué qu'elle pouvait avoir quatre racines. La voie est alors ouverte plus largement encore avec les travaux de Viète sur les relations entre coefficients et racines d'une équation algébrique, puis ceux de Girard (1595-1632) qui prolonge l'idée — énoncée par Cardan — de l'ordre de multiplicité d'une racine, et va aboutir au célèbre *théorème de Gauss-d'Alembert*.

La méthode de Ferrari transforme l'équation de départ, du quatrième degré, en une équation du troisième degré, théoriquement résoluble. Comme pour le second degré, on considère le premier membre de **(4)** comme le début du développement d'un carré parfait dont un terme est inconnu, et on a :

$$\left(x^2 + \frac{p}{2}x + t\right)^2 - \left[\left(t + \frac{p^2}{4} - q\right)x^2 + \left(\frac{pt}{2} - r\right)x + \frac{t^2}{4} - s\right] = 0$$

Il suffit alors de chercher à déterminer t de telle sorte que le crochet soit un carré parfait, pour que l'on n'ait plus à résoudre que deux équations simples du second degré. Cette condition est réalisée dès que le trinôme entre crochets admet une racine double, soit si :

$$\left(\frac{pt}{2} - r\right)^2 - 4\left(t + \frac{p^2}{4} - q\right)\left(\frac{t^2}{4} - s\right) = 0$$

et l'on reconnait là une équation du 3e degré en t.

La résolution des équations de degré supérieur à quatre posa dès lors un problème à tous les mathématiciens qui ont longtemps essayé de poursuivre, avec le 5e degré, les méthodes de résolution par radicaux. L'obstacle fut infranchissable, et, vers la fin du XVIIIe siècle, Leibnitz est encore l'un des derniers à se préoccuper de ce problème. Une longue période de tâtonnements s'ouvrait alors, et ce sont les recherches de Lagrange (1736-1813) qui vont dégager, à partir des relations entre coefficients et racines, la voie que Galois (1811-1832) va prendre pour résoudre d'une façon globale le problème des équations algébriques.

Polynômes

On désigne par K un corps commutatif, et par K [X] l'anneau des polynômes à une indéterminée à coefficients dans K, c'est-à-dire des expressions de la forme $\sum_{n=0}^{\infty} a_n X^n$ où tous les scalaires a_n du corps K sont nuls à partir d'un certain rang. On rappelle qu'un idéal I d'un anneau A est un sous-groupe additif de A tel que, si $X \in A$ et $y \in I$, alors $Xy \in I$. Par conséquent, tout idéal est un sous-anneau (la réciproque étant fausse). Dans l'anneau K [X], les idéaux possèdent des propriétés très particulières, dont la plus importante est sans nul doute d'être principaux : c'est-à-dire que pour tout idéal $\mathfrak{J} \subset K[X]$, on peut trouver un polynôme $P \in \mathfrak{J}$ tel que tout élément de $\mathfrak{J}$ soit un multiple de P par un élément de K [X] ; donc pour tout $Q \in \mathfrak{J}$, il existe un polynôme $A \in K[X]$ tel que $Q = P \cdot A$. Ce polynôme P est unique à une constante multiplicative près. Celui dont le coefficient du terme de plus haut degré est 1 est dit base de l'idéal.

On dit qu'un idéal $\mathfrak{J}$ est premier si la condition :

$$A \cdot B \in \mathfrak{J} \Rightarrow A \in \mathfrak{J} \quad \text{ou bien } B \in \mathfrak{J}.$$

On peut rapprocher cette définition de celle d'un polynôme p (X) irréductible si :

$$p(X) = q(X)\, r(X) \Rightarrow q(X) = c \in K \quad \text{ou bien } r(X) = c \in K.$$

On montre d'ailleurs que si un idéal $\mathfrak{J}$ est premier, son polynôme de base est irréductible, et réciproquement.

Soit L un corps contenant le corps K, on dit que L est un sur-corps de K ; par exemple, $\mathbb{C}$ est un sur-corps de $\mathbb{R}$ qui est lui-même un sur-corps de $\mathbb{Q}$. La loi externe définie par : $(l, k) \to l \cdot k$ où $l \in L$ et $k \in K$, donne à L une structure d'algèbre sur K ; on dit alors que L est une extension de K, par exemple $\mathbb{C}$ est une extension de $\mathbb{R}$. Soit alors $\alpha \in L$, à tout polynôme $f(X) \in K[X]$, on peut faire correspondre le scalaire $f(\alpha) \in L$; et lorsque f (X) décrit K [X], les $f(\alpha)$ correspondants décrivent un sous-anneau de L, dit engendré par K et α, et que l'on note $K[\alpha]$.

En reprenant la même construction, mais cette fois avec le corps K (X) des fractions rationnelles en X à coefficients dans K, quotient de deux polynômes de K [X], soit $\frac{f(X)}{g(X)}$, si $\alpha \in L$, on obtient un sous-corps de L en considérant les scalaires $f(\alpha)/g(\alpha)$ pour tous les polynômes $f(X) \in K[X]$ et $g(X) \in K[X]$ tels que $g(\alpha) \neq 0$; on l'appelle sous-corps engendré par K et α et on le note $K(\alpha)$. On remarque tout de suite que si $\alpha \in K$, alors $K(\alpha) = K$, et si $\alpha \notin K$, $K(\alpha) \underset{\neq}{\supset} K$.

Définition

Le sur-corps L de K est une extension simple de K, s'il existe un élément $\alpha \in L$ tel que $L = K(\alpha)$.

Ce sont ces extensions qui permettent de rechercher l'ensemble des racines d'une équation algébrique : on considère un polynôme f (X) irréductible sur un corps K, et on va chercher un sur-corps de K dans lequel f (X) possède une racine.

Définition

Soit $L = K(\alpha)$ une extension simple de K. Elle est transcendante sur K s'il n'existe aucun polynôme f (X) non nul, à coefficients dans K, tel que $f(\alpha) = 0$. Elle est dite algébrique dans le cas contraire. Ce sont évidemment les extensions algébriques simples qui vont permettre l'étude des équations algébriques. Leurs propriétés sont très larges.

On montre que, si $K(\alpha)$ est algébrique simple, alors le corps $K(\alpha)$ est égal à l'anneau $K[\alpha]$. Ceci permet de voir que toute extension algébrique simple $K(\alpha)$ est isomorphe à l'anneau quotient de K [X] par l'idéal des polynômes multiples d'un certain polynôme irréductible normé (c'est-à-dire dont le coefficient du terme de plus haut degré est égal à l'unité). Soit d le degré de ce polynôme irréductible ; alors tout élément z de $k(\alpha)$ se met de façon unique sous la forme :

$$z = a_0 + a_1\alpha + \ldots\ldots + a_{d-1}\alpha^{d-1}$$

où $a_i \in K$ pour tout i.

En effet, puisque $z \in K(\alpha)$, $z \in K[\alpha]$, donc on peut trouver un polynôme f tel que $z = f(\alpha)$. Soit p (X) le polynôme irréductible de degré d. La division de f (X) par p (X) s'écrit : $f(X) = q(X)\, p(X) + r(X)$ avec r (X) polynôme de degré inférieur à d, ou bien $r(X) = 0$. Par conséquent, $f(\alpha) = r(\alpha)$ puisque $p(\alpha) = 0$. On a donc bien une représentation du type indiqué pour z.

On reconnaît donc que $K(\alpha)$ est un espace vectoriel sur K, de dimension finie, égale à d. Ce nombre s'appelle le degré de l'extension $K(\alpha)$ et se note $d = [K(\alpha) : K]$. Les extensions transcendantes simples de K sont, par contre, des espaces vectoriels de dimension infinie sur K.

Le corps des nombres complexes $\mathbb{C}$ est une extension de degré 2 (ou encore quadratique), soit $\mathbb{R}(i)$ du corps des réels $\mathbb{R}$, le polynôme irréductible associé étant $x^2 + 1$; et l'on sait bien que les éléments de $\mathbb{C}$ s'écrivent de façon unique $a + bi$. Le corps $\mathbb{Q}(\sqrt{2})$ dont les éléments se représentent par $a + b\sqrt{2}$ (où $a \in \mathbb{Q}$ et $b \in \mathbb{Q}$), est une extension quadratique du corps $\mathbb{Q}$ des nombres rationnels, le polynôme irréductible associé étant $x^2 - 2$.

Si maintenant on se donne un polynôme $p(X) \in K[X]$ irréductible sur K, on cherchera donc une extension algébrique simple $K(\alpha)$ dont le polynôme irréductible associé soit p (X), afin de trouver une racine de p (X) dans ce sur-corps.

L'étude qui vient d'être faite montre que le corps $K[X]/(p(X))$ satisfait aux conditions imposées, en désignant par $(p(X))$ l'idéal des polynômes multiples de p (X). Cette construction permet de définir un symbole α, racine d'un polynôme irréductible sur K. On peut noter que l'extension cherchée est unique à un isomorphisme près laissant les éléments de K invariants (on dit un K-isomorphisme).

On donne à cette construction le nom justifié d'*adjonction symbolique*. C'est ainsi que Cauchy construisit les nombres complexes par le corps quotient

$$\mathbb{R}[X]/(X^2 + 1).$$

Il faut bien sûr rapprocher ces notions d'extensions algébriques ou transcendantes de celles de nombres algébriques ou transcendants sur un corps K, le parallélisme entre les définitions étant évident.

L'adjonction symbolique que nous venons de définir montre donc que, si l'on se donne un corps K, un polynôme p (X) irréductible dans K, alors il existe une extension algébrique simple (de degré fini) L telle que, si p (X) possède une racine dans L, il en possède m (si m est le degré de p (X)) ; autrement dit, telle que p (X) s'y décompose en facteurs irréductibles de degré 1. Une telle extension est dite galoisienne.

Il s'agit là de l'une des notions de base posées par Galois pour sa théorie des équations algébriques. C'est Lagrange qui lui avait ouvert la voie en observant sur les exemples d'équations de degrés 3 et 4 le grand intérêt de l'étude du nombre N de valeurs prises par une fonction rationnelle des racines lorsque l'on permute celles-ci. Il définit les « résolvantes de Lagrange » :

$$z_p = \sum_{k=1}^{n} \omega_p^k x_k,$$

où $x_1, x_2, \ldots, x_n$ désignent les n racines d'une équation algébrique de degré n et $(\omega_p)_{p=1,\ldots,n}$ les racines n-ièmes de l'unité, et montre dans un très célèbre mémoire que la détermination des racines x_k est assurée dès que l'on connaît les nombres z_p. Un mémoire de Vandermonde (1735-1796), mathématicien français, dont les résultats ne furent entièrement démontrés que par Carl Friedrich Gauss (1777-1855), complète les recherches de Lagrange.

▶ *Le mathématicien norvégien* Niels-Henrik Abel *(1802-1829) ; il montra que les équations algébriques générales ne peuvent pas être résolues algébriquement quand leur degré est supérieur au quatrième.*



Groupes de Galois — Théorème d'Abel

On rappelle qu'une permutation d'un ensemble E est une bijection de E sur E. Lorsque E est un ensemble fini, on représente une permutation par un tableau à deux lignes, chaque élément de la ligne supérieure étant placé au-dessus de son image. L'ensemble des permutations d'un ensemble fini, muni de la loi de composition usuelle des applications, forme un groupe, que l'on appelle le groupe symétrique.

Soit K un corps et L une extension galoisienne de K. Les automorphismes de L qui laissent K invariant (automorphies par rapport à K, de L) forment un groupe, qui est le groupe de Galois de L par rapport à K. C'est à Julius Dedekind (1831-1916) que l'on doit cette définition remplaçant celle du groupe de permutations des racines d'une équation telle que Galois, en 1830, l'avait énoncée. Le lien entre ces deux formes s'établit comme suit : si p (X) est un polynôme irréductible dans un corps K, les racines $x_1, x_2, ..., x_p$ de ce polynôme engendrent une extension galoisienne L de K. Soit G le groupe de Galois de L par rapport à K ; si f est un élément de G, toute racine de p (X) est transformée par f en une autre racine. Par conséquent, f induit sur les racines de p (X) une permutation qui, réciproquement, détermine f.

Soit alors q (X) un autre polynôme irréductible dans K et M l'extension de K construite à partir des zéros $y_1, y_2, ..., y_q$ de q (X). On voit alors que, lorsque la résolution de p (X) = 0 donne celle de q (X) = 0, les racines de q (X) sont des expressions rationnelles de celles de p (X), donc sont des éléments de L ; par suite, M est un sous-corps de L. Le principe de Galois consiste à construire ces corps intermédiaires échelonnés entre K et L. Soit T un de ces corps, alors les T-automorphismes de L forment un sous-groupe H de G. Réciproquement, soit H un sous-groupe de G ; les éléments de L invariants par les automorphismes de H forment un sur-corps S de K. On a donc construit une correspondance entre les corps intermédiaires entre K et L et les sous-groupes de G, et une autre en sens contraire. C'est là que le théorème fondamental de Galois se place en montrant que ces deux correspondances sont réciproques. C'est-à-dire que, si T est un corps intermédiaire et si H est le groupe des T-automorphismes de L, alors T est aussi le sous-corps de L formé par les éléments invariants par tous les éléments de H. Donc, si H est un sous-groupe de G et si T est le corps formé par les éléments invariants par les éléments de H, alors H est aussi le groupe des T-automorphismes de L. Pour cette raison, on dit que le corps T et le groupe H sont associés.

La théorie de Galois permet de résoudre le problème de la résolubilité des équations radicaux, qui consiste à trouver une expression formée des quatre opérations et de l'extraction de racines p-ièmes qui soit solution d'une équation p (X) = 0, algébrique. On a pu ainsi trouver la condition générale pour qu'une équation soit résoluble par radicaux.

Soit donc un corps K et un polynôme irréductible dans K, p (X). Chaque racine de p (X) construite par radicaux est algébrique sur K : elle est racine d'un polynôme q (X) irréductible dans K. Soit L l'extension de K construite à partir de ces radicaux (racines de p (X)) et de tous les autres zéros des polynômes q (X) ; soit G le groupe de Galois de cette extension, qui est galoisienne. Les racines de p (X) sont dans L, et engendrent un corps M intermédiaire entre K et L. On sait qu'on peut associer à M un sous-groupe H de G, et que le groupe de Galois de l'extension L est alors G/H.

Le mathématicien norvégien Niels-Henrik Abel (1802-1829) a démontré que les groupes de Galois engendrés par des radicaux possèdent la propriété dite de résolubilité ; c'est-à-dire qu'il existe une suite de sous-groupes G_1 (= G), $G_2, ..., G_{n-1}, G_n$ (= {1}) commençant par G, finissant par le groupe formé du seul élément unité (dit groupe unité), tels que chaque G_i soit un sous-groupe distingué de G_{i-1} (c'est-à-dire $G_i G_{i-1} G_i^{-1} \subset G_i$), le groupe quotient G_{i-1}/G_i étant cyclique (c'est-à-dire fini et engendré par un seul élément). On montre ainsi que le groupe des permutations d'un ensemble à 3 éléments est résoluble ; de même le groupe — à 24 éléments — des permutations d'un ensemble à 4 éléments est résoluble : il est engendré en fait par les permutations :

$$\pi_1 = \begin{pmatrix} a\ b\ c\ d \\ b\ a\ d\ c \end{pmatrix}, \quad \pi_2 = \begin{pmatrix} a\ b\ c\ d \\ b\ c\ a\ d \end{pmatrix}, \quad \pi_3 = \begin{pmatrix} a\ b\ c\ d \\ b\ a\ c\ d \end{pmatrix}$$

et si l'on pose G_2 = groupe engendré par π_1 et π_2, G_3 = groupe engendré par π_1 et $\pi_2 \pi_1 \pi_3^{-1}$,

$$\left[\text{on rappelle que } \pi_3^{-1} = \begin{pmatrix} a\ b\ c\ d \\ b\ a\ c\ d \end{pmatrix}\right]$$

et G_4 = groupe engendré par π_1, alors la suite G, G_2, G_3, G_4, $\mathcal{E}$ où $\mathcal{E}$ est formé par la seule $\pi_0 = \begin{pmatrix} a\ b\ c\ d \\ a\ b\ c\ d \end{pmatrix}$ vérifie les conditions énoncées, les groupes quotients ayant pour ordres respectifs 2, 3, 2, 2.

Si maintenant on en revient au problème de départ, on voit (théorème d'Abel) que, pour qu'une équation algébrique soit résoluble par radicaux, il faut et il suffit que son groupe de Galois soit résoluble.

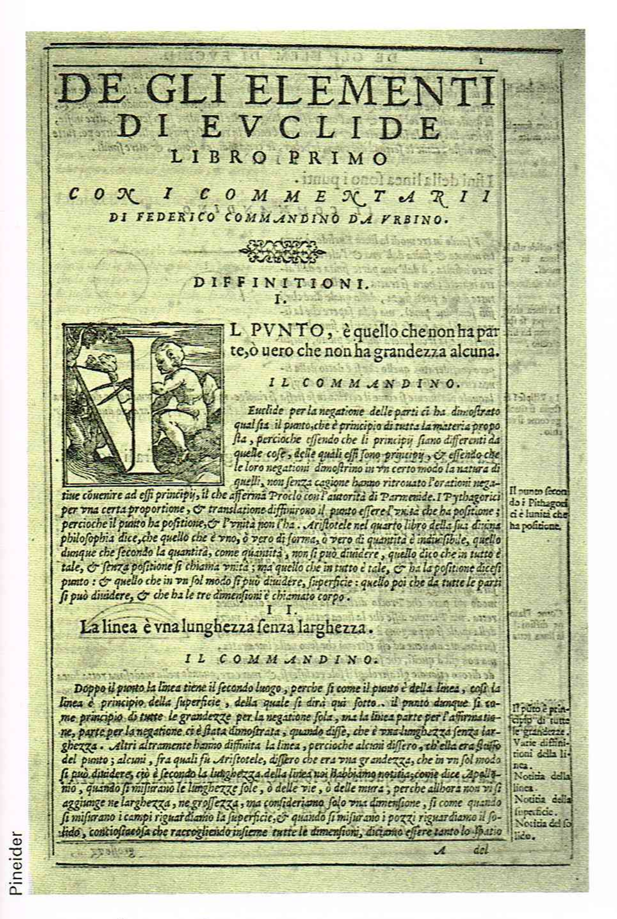
Il est donc maintenant évident que toute équation de degré 3 est résoluble par radicaux (précédemment, on a donné les formules explicites de Cardan), et qu'il en est de même pour toutes les équations de degré 4.

L'étude du groupe de permutation d'un ensemble à n éléments, pour $n \geqslant 5$, que l'on ne peut faire ici en détail, montre qu'il n'est en général pas résoluble (à moins d'imposer des conditions restrictives sur l'équation). Le théorème d'Abel montre donc qu'une équation de degré supérieur ou égal à 5 ne peut en général être résoluble par radicaux.

Ce théorème met donc ainsi un point final à ce très vieux problème pour lequel les plus grands mathématiciens ont essayé les plus diverses des méthodes pour trouver une solution. C'est d'ailleurs en essayant de résoudre par radicaux l'équation de degré 5 qu'Abel se convainquit de l'impossibilité de le faire et, reprenant la voie ouverte par Lagrange — sur les permutations des racines dans leurs expressions rationnelles — énonça son résultat en utilisant l'extraordinaire théorie de Galois.

BIBLIOGRAPHIE

DUBREUIL-JACOTIN P. et M., *Algèbre,* Dunod. - GALOIS É., *Œuvres,* Gauthier-Villars. - LANG S., *Algèbre.* - SAMUEL P., *Théorie algébrique des nombres,* Hermann.



1

DE GLI ELEMENTI DI EVCLIDE

LIBRO PRIMO

CON I COMMENTARII DI FEDERICO COMMANDINO DA VRBINO.

DIFFINITIONI.

I.

IL PVNTO, è quello che non ha parte, ò uero che non ha grandezza alcuna.

IL COMMANDINO.

Euclide per la negatione delle parti ci ha dimoſtrato qual ſia il punto, che è principio di tutta la materia propoſta, percioche eſſendo che li principij ſiano differenti da quelle coſe, delle quali eſſi ſono principij, & eſſendo che le loro negationi dimoſtrino in vn certo modo la natura di quelli, non ſenza cagione hanno ritrouato l'orationi negatiue cõuenire ad eſſi principij, il che afferma Proclo con l'autorità di Parmenide. I Pythagorici per vna certa proportione, & translatione diffinirono il punto eſſere l'vnità che ha poſitione; percioche il punto ha poſitione, & l'vnità non l'ha. Ariſtotele nel quarto libro della ſua diuina philoſophia dice, che quello che è vno, ò vero di forma, ò vero di quantità è indiuiſibile, quello dunque che ſecondo la quantità, come quantità, non ſi può diuidere, quello dico che in tutto è tale, & ſenza poſitione ſi chiama vnità; ma quello che in tutto è tale, & ha la poſitione diceſi punto: & quello che in vn ſol modo ſi può diuidere, ſuperficie: quello poi che da tutte le parti ſi può diuidere, & che ha le tre dimenſioni è chiamato corpo.

Il punto ſecondo i Pithagorici è l'unità che ha poſitione.

II.

La linea è vna lunghezza ſenza larghezza.

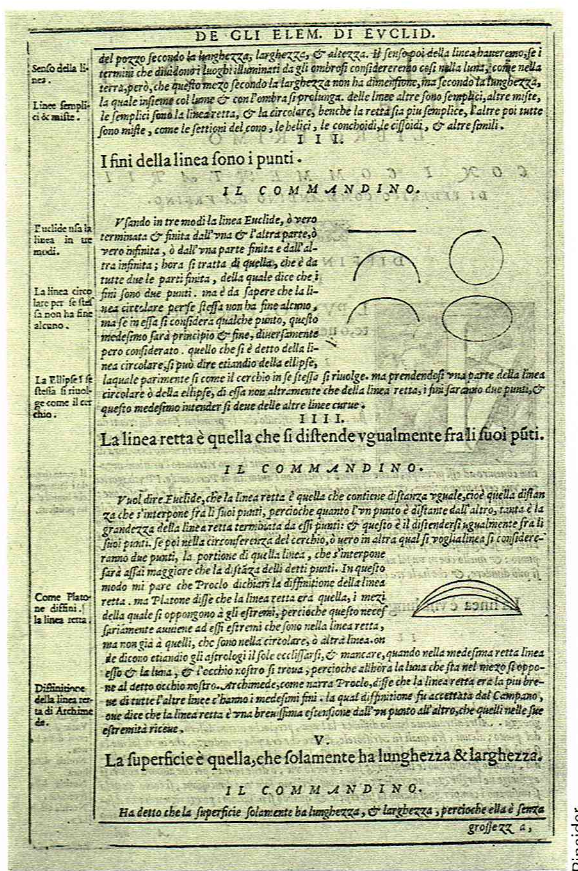
IL COMMANDINO.

Doppo il punto la linea tiene il ſecondo luogo, perche ſi come il punto è della linea, coſì la linea è principio della ſuperficie, della quale ſi dirà qui ſotto. il punto dunque ſi come principio di tutte le grandezze per la negatione ſola, ma la linea parte per l'affirmatione, parte per la negatione ci è ſtata dimoſtrata, quando diſſe, che è vna lunghezza ſenza larghezza. Altri altramente hanno diffinita la linea, percioche alcuni diſſero, ch'ella era fluſſo del punto; alcuni, fra quali fu Ariſtotele, diſſero che era vna grandezza, che in vn ſol modo ſi può diuidere, ciò è ſecondo la lunghezza. della linea noi habbiamo notitia, come dice Apollonio, quando ſi miſurano le lunghezze ſole, ò delle vie, ò delle mura, perche allhora non vi ſi aggiunge ne larghezza, ne groſſezza, ma conſideriamo ſolo vna dimenſione, ſi come quando ſi miſurano i campi riguardiamo la ſuperficie, & quando ſi miſurano i pozzi riguardiamo il ſolido, conciosiacoſa che raccogliendo inſieme tutte le dimenſioni, diciamo eſſere tanto lo ſpatio

Il puto è principio di tutte le grandezze. Varie diffinitioni della linea. Notitia della linea. Notitia della ſuperficie. Notitia del ſolido.

A del

Pineider



DE GLI ELEM. DI EVCLID.

del pozzo ſecondo la lunghezza, larghezza, & altezza. il ſenſo poi della linea haueremo, ſe i termini che diuidono i luoghi illuminati da gli ombroſi conſidereremo coſi nella luna, come nella terra, però, che queſto mezo ſecondo la larghezza non ha dimenſione, ma ſecondo la lunghezza, la quale inſieme col lume & con l'ombra ſi prolunga. delle linee altre ſono ſemplici, altre miſte, le ſemplici ſono la linea retta, & la circolare, benche la retta ſia piu ſemplice, l'altre poi tutte ſono miſte, come le ſettioni del cono, le helici, le conchoidi, le ciſſoidi, & altre ſimili.

Senſo della linea. Linee ſemplici & miſte.

III.

I fini della linea ſono i punti.

IL COMMANDINO.

Vſando in tre modi la linea Euclide, ò vero terminata & finita dall'vna & l'altra parte, ò vero infinita, ò dall'vna parte finita e dall'altra infinita; hora ſi tratta di quella, che è da tutte due le parti finita, della quale dice che i fini ſono due punti. ma è da ſapere che la linea circolare per ſe ſteſſa non ha fine alcuno, ma ſe in eſſa ſi conſidera qualche punto, queſto medeſimo ſarà principio & fine, diuerſamente pero conſiderato. quello che ſi è detto della linea circolare, ſi può dire etiandio della ellipſe, laquale parimente ſi come il cerchio in ſe ſteſſa ſi riuolge. ma prendendoſi vna parte della linea circolare ò della ellipſe, di eſſa non altramente che della linea retta, i fini ſaranno due punti, & queſto medeſimo intender ſi deue delle altre linee curue.

Euclide uſa la linea in tre modi. La linea circolare per ſe ſteſſa non ha fine alcuno. La Ellipſe I ſe ſteſſa ſi riuolge come il cerchio.

IIII.

La linea retta è quella che ſi diſtende vgualmente fra li ſuoi pũti.

IL COMMANDINO.

Vuol dire Euclide, che la linea retta è quella che contiene diſtanza vguale, cioè quella diſtanza che s'interpone fra li ſuoi punti, percioche quanto l'vn punto è diſtante dall'altro, tanta è la grandezza della linea retta terminata da eſſi punti: & queſto è il diſtenderſi ugualmente fra li ſuoi punti. ſe poi nella circonferenza del cerchio, ò uero in altra qual ſi voglia linea ſi conſidereranno due punti, la portione di quella linea, che s'interpone ſarà aſſai maggiore che la diſtãza delli detti punti. In queſto modo mi pare che Proclo dichiari la diffinitione della linea retta. ma Platone diſſe che la linea retta era quella, i mezi della quale ſi oppongono à gli eſtremi, percioche queſto neceſſariamente auuiene ad eſſi eſtremi che ſono nella linea retta, ma non già à quelli, che ſono nella circolare, ò altra linea. onde dicono etiandio gli aſtrologi il ſole eccliſſarſi, & mancare, quando nella medeſima retta linea eſſo & la luna, & l'occhio noſtro ſi troua, percioche allhora la luna che ſta nel mezo ſi oppone al detto occhio noſtro. Archimede, come narra Proclo, diſſe che la linea retta era la piu breue di tutte l'altre linee c'hanno i medeſimi fini. la qual diffinitione fu accettata dal Campano, oue dice che la linea retta è vna breuiſſima eſtenſione dall'vn punto all'altro, che quelli nelle ſue eſtremità riceue.

Come Platone diffini la linea retta. Diffinitione della linea retta di Archimede.

V.

La ſuperficie è quella, che ſolamente ha lunghezza & larghezza.

IL COMMANDINO.

Ha detto che la ſuperficie ſolamente ha lunghezza, & larghezza, percioche ella è ſenza

groſſezz a,

Pineider

◀ ***Deux pages extraites des Éléments d'Euclide, œuvre qui peut être considérée comme l'une des premières synthèses de la géométrie.***

LA GÉOMÉTRIE

La *géométrie,* comme son nom l'indique, naquit du besoin pratique de mesurer des parcelles de terrain. Il est certain que déjà les Égyptiens avaient trouvé des méthodes de mesure et avaient même élaboré un début de science géométrique. La tradition veut que la géométrie grecque naisse vers le VIe siècle av. J.-C. avec *Thalès de Milet*. Elle s'est ensuite progressivement affinée et développée avec *Pythagore,* au Ve siècle, *Eudoxe,* au IVe siècle, qui obtinrent un nombre appréciable de résultats géométriques : inscription de sphères dans un cône, similitude des triangles, principales propriétés du cercle, polygones et polyèdres réguliers, sections coniques. En utilisant et en complétant ces résultats, *Euclide* (fin du IVe siècle av. J.-C.) réalisa avec ses *Éléments* la première synthèse de la géométrie. Il eut le souci de fonder la géométrie et donna un exemple important d'élaboration d'un système axiomatique. Aujourd'hui, les fondements de ce système se révèlent très peu assurés. C'est seulement à la fin du XIXe siècle — et surtout avec *D. Hilbert* — que la géométrie euclidienne fut fondée de façon satisfaisante.

D. Hilbert, dans son livre *Grundlagen der Geometrie* (1899), donna la première construction axiomatique qui ne devait rien à l'intuition. Le système d'axiomes utilisé est divisé en cinq groupes; la grande originalité de son œuvre est que le contenu concret, ou conceptuel, n'est pas présupposé, mais est introduit par la formulation des axiomes. D'autre part, D. Hilbert formula les conditions auxquelles doit obéir un système d'axiomes pour être satisfaisant : cohérence, indépendance, saturation. Il établit l'indépendance du postulat d'Euclide sur les parallèles par rapport aux autres axiomes et prouva aussi qu'on peut construire des géométries non contradictoires en conservant tous les axiomes euclidiens, sauf celui des parallèles.

Aujourd'hui, la géométrie comporte un domaine de recherche immense, allant de la géométrie plane aux géométries à *n* dimensions, de la géométrie classique aux géométries définies par un groupe de transformations (selon le *Programme d'Erlangen* de *F. Klein*), de la géométrie algébrique à la géométrie différentielle et à l'*analysis situs* ou topologie.

La géométrie élémentaire

La géométrie plane

Les axiomes

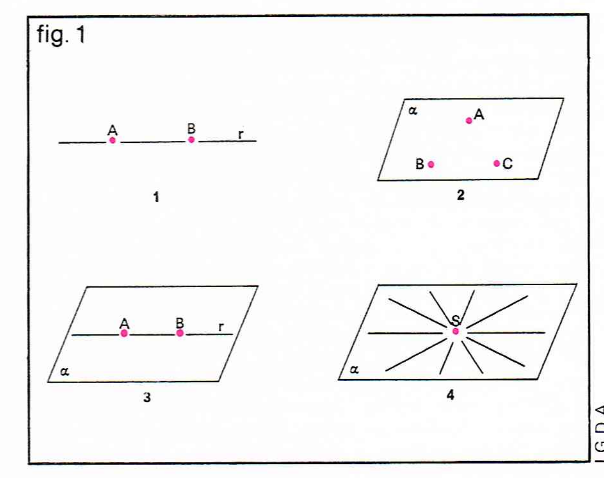
En géométrie plane, on considère un ensemble de base appelé *plan*. Les éléments de cet ensemble sont appelés *points*. Une partie quelconque d'un plan se nomme une *figure*.

D. Hilbert parvint le premier à formuler une trentaine d'axiomes qui, non seulement, introduisent des relations entre le plan et certaines de ses figures, mais encore construisent le plan et définissent les figures fondamentales.

Ces axiomes se divisent en cinq groupes :

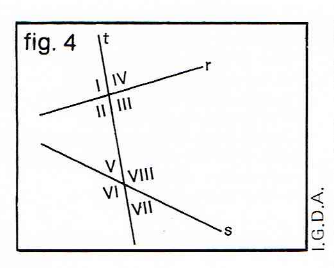
— axiomes d'appartenance,
— axiomes d'ordre,
— axiomes de congruence ou d'égalité,
— axiome des parallèles (ou d'Euclide),
— axiomes de continuité.

● *Les axiomes d'appartenance* expriment les relations d'appartenance et d'inclusion entre points, droites et plans; par exemple : « par deux points passe une droite et une seule » (*fig.* 1).

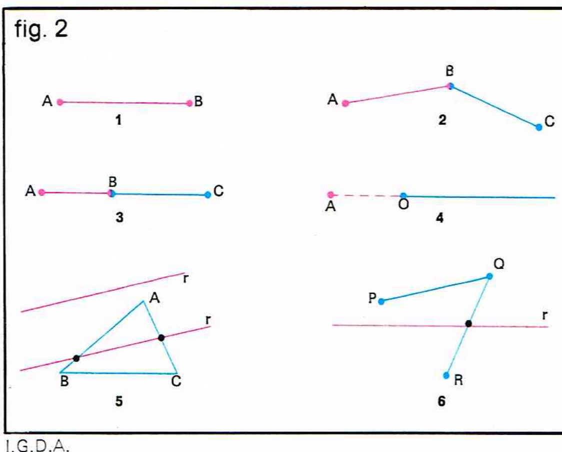


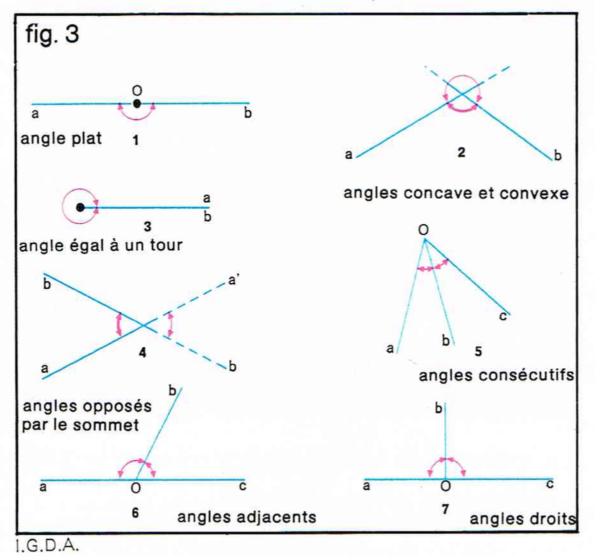
I.G.D.A.

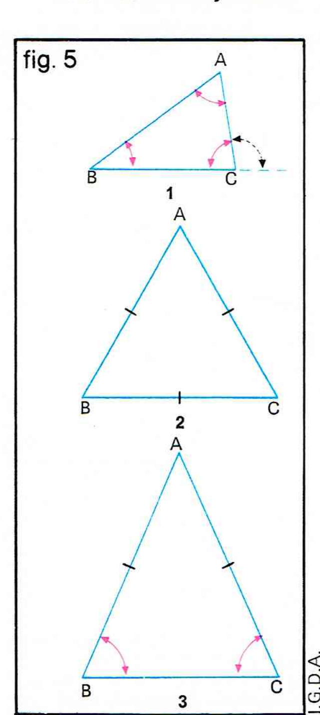
◀ **Figure 1 : 1)** ***deux points distincts déterminent une droite; 2) trois points distincts n'appartenant pas à une même droite déterminent un plan; 3) si deux points d'une droite appartiennent à un plan, alors la droite appartient à ce plan; 4) l'ensemble des droites d'un même plan sécantes en un point S, forme un faisceau de centre S.***

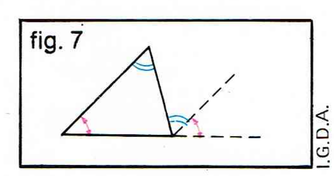


▲▶ *A gauche,* **figure 4** — *une droite coupant deux droites coplanaires, détermine des angles : alternes-externes (I-VII et IV-VI) ; alternes-internes (II-VIII et III-V) ; correspondants (I-V, II-VI, IV-VIII et III-VII) ; internes conjugués (II-V et III-VIII) ; externes conjugués (I-VI et IV-VII).*
A droite, **figure 2** : 1) *le segment AB ;* 2) *deux segments consécutifs ;* 3) *deux segments adjacents ;* 4) *une demi-droite d'origine O ;* 5) *axiome de Pasch ;* 6) P *et* Q *appartiennent au même demi-plan de bord* r, Q *et* R *appartiennent chacun à un des demi-plans de bord* r *et il existe alors un point* r *entre* Q *et* R.
▼ *Ci-dessous,* **figure 5** — *triangles :* 1) *angle externe à un angle d'un triangle ;* 2) *triangle équilatéral ;* 3) *triangle isocèle.*
En bas, **figure 7** : *un angle externe d'un triangle est égal à la somme des angles internes non adjacents.*









● *Les axiomes d'ordre* précisent l'emploi et les propriétés du mot « entre » ; par exemple : « sur la droite définie par deux points distincts A et B, il existe au moins un point C situé entre ces deux points (et donc au moins une infinité dénombrable de points) » ; l'axiome de *Pasch* affirme que : « si A, B et C sont trois points non alignés, et *r*, une droite de leur plan telle qu'aucun des trois points ne lui appartienne, alors *r* coupe deux ou aucun des segments parmi les segments AB, BC et AC » (le segment AB est constitué, par définition, de A et de B et des points situés entre A et B) (*fig.* 2).

● *Les axiomes d'égalité* expriment les propriétés de l'égalité géométrique, sans recours à l'intuition du déplacement des objets indéformables. L'égalité géométrique ne doit pas être confondue avec la notion d'égalité de la logique et de la théorie des ensembles : deux segments distincts peuvent être égaux, pourvu qu'ils appartiennent à la même classe d'équivalence (l'égalité étant ici une relation d'équivalence) ; on parlera plutôt par la suite de *figures isométriques* que de « figures égales » ; ce qualificatif, qui évoque une *égalité de mesure,* sera justifié plus loin. En considérant une classe de segments deux à deux égaux, on arrive au concept de *longueur,* et en considérant une classe d'angles deux à deux égaux, on arrive au concept d'*amplitude angulaire* (*fig.* 3).

● *L'axiome des parallèles.* On peut tirer des axiomes précédents que, par un point A extérieur à une droite *r*, il est possible de mener une droite qui ne coupe pas *r*. Par définition, deux droites parallèles sont deux droites coplanaires qui ne se coupent pas.

Axiome d'Euclide : soit une droite *r* et A un point extérieur à *r*, dans le plan déterminé par *r* et A, il existe, au plus, une droite qui passe par A et qui ne coupe pas *r*.

Il résulte de ce qui précède et de ce dernier axiome que, par un point extérieur à une droite, il passe une unique parallèle à cette droite. L'introduction de cet axiome des parallèles simplifie les fondements de la géométrie et allège considérablement son élaboration.

On démontre notamment que deux droites coplanaires formant avec une autre droite qui les coupe des angles alternes-internes égaux sont parallèles ; ou encore : que deux droites coplanaires formant avec une autre droite qui les coupe des angles conjugués supplémentaires (leur somme est égale à un angle plat) sont parallèles ; ou encore des angles correspondants, ou alternes-externes, égaux (*fig.* 4).

On démontre aussi que des segments parallèles qui ont leurs extrémités sur deux droites parallèles sont égaux ; que si deux droites sont parallèles, toute droite perpendiculaire à l'une est perpendiculaire à l'autre.

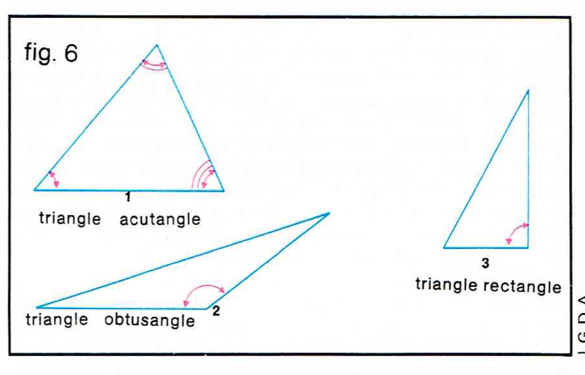
● *Les axiomes de la continuité* précisent enfin les conditions de passage à la limite dans l'espace et de mesure des grandeurs. L'*axiome d'Archimède* permet l'introduction de la continuité en géométrie : « Entre deux points A, B d'une droite, choisissons un point A_1 et déterminons des points distincts A_2, A_3, ..., tels que les segments AA_1, A_1A_2, A_2A_3, ... soient égaux ; il existera toujours un point A_n tel que B soit situé entre A et A_n. » On peut écrire aussi l'axiome d'Archimède pour les angles. Les axiomes et postulats euclidiens ont été reconnus insuffisants pour justifier les raisonnements des « éléments » fondés sur la continuité et sur l'ordre ; ce qui amena *Dedekind, Cantor, Weierstrass,* vers 1870, à formuler un *postulat de la continuité de la droite,* et d'autre part *Pasch* à introduire le *postulat d'ordre* à travers les axiomes de la géométrie.

Correspondance biunivoque – Figures isométriques

Par définition, deux figures sont dites *isométriques* (ou égales) s'il existe entre elles une correspondance ponctuelle biunivoque, appelée aussi *isométrie,* telle que les segments et les angles déterminés par les points qui se correspondent soient toujours égaux. Si les correspondants de trois points A, B, C sont respectivement A', B', C', on appelle correspondant ou *homologue* à AB le segment A'B', et ainsi de suite.

Le triangle

Par définition, on appelle *triangle* ABC (si A, B et C sont trois points non alignés) la figure constituée des segments AB, BC, AC appelés côtés et des points simultanément intérieurs aux angles convexes $\widehat{ABC}$, $\widehat{BCA}$, $\widehat{CAB}$ (*fig.* 5).



Remarquons que tout triangle est une figure *convexe,* une figure *convexe* étant par définition une figure telle que, si deux points A et B lui appartiennent, alors le segment AB lui appartient. Il y a trois cas bien connus d'égalité des triangles :

— deux triangles ayant deux côtés égaux et l'angle compris entre ces côtés égal sont égaux ;

— deux triangles ayant un côté égal et les angles adjacents à ce côté égaux sont égaux ;

— deux triangles ayant leurs trois côtés égaux sont égaux.

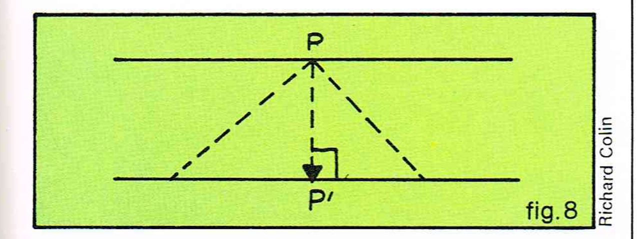
On démontre qu'un côté d'un triangle est plus petit que la somme des deux autres et plus grand que leur différence. Ce résultat est généralisé en topologie sous le nom d'inégalité triangulaire (voir *Topologie*) (*fig.* 6).

On démontre que le postulat d'Euclide implique que la somme des angles d'un triangle plan est égale à un angle plat ou, de même, que tout angle externe d'un triangle est égal à la somme des angles internes non adjacents (*fig.* 7).

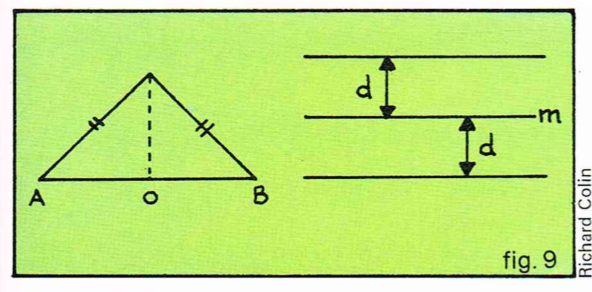
Distances, lieux, symétries

Par définition, la *distance* entre deux points est la longueur de leur segment ; la distance d'un point à une

droite est la distance de ce point à sa projection sur la droite. On en déduit que la distance d'un point à une droite est le minimum des longueurs des segments ayant pour extrémités le point considéré et un point de la droite. On définit la distance de deux droites parallèles comme la distance d'un point de l'une à l'autre (*fig.* 8).



On appelle *lieu des points* qui jouissent d'une propriété donnée l'ensemble de tous les points pour lesquels cette propriété est vérifiée. On démontre facilement que le lieu des points équidistants des extrémités d'un segment est la perpendiculaire élevée en son centre qu'on appelle *axe* du segment. On démontre aussi que le lieu des points d'un plan à une distance donnée *d* d'une de ses droites *m* est la réunion de deux parallèles à la droite (*fig.* 9).



Par définition, une circonférence (ou par extension, un cercle) est le lieu des points d'un plan à égale distance d'un point, appelé *centre*. Cette distance commune se nomme *rayon*, et le *diamètre* est le double du rayon (*fig.* 10 et 11).

On démontre que deux cercles dont les rayons sont égaux sont égaux.

Par définition, deux points sont dits *symétriques* par rapport à un point O si O est le milieu de leur segment; on les dit symétriques par rapport à une droite *s* (axe de symétrie) si celle-ci est l'axe de leur segment. La figure F' symétrique d'une figure F par rapport à O (respectivement par rapport à *s*) est lieu des points symétriques des points de F par rapport à O (respectivement par rapport à *s*).

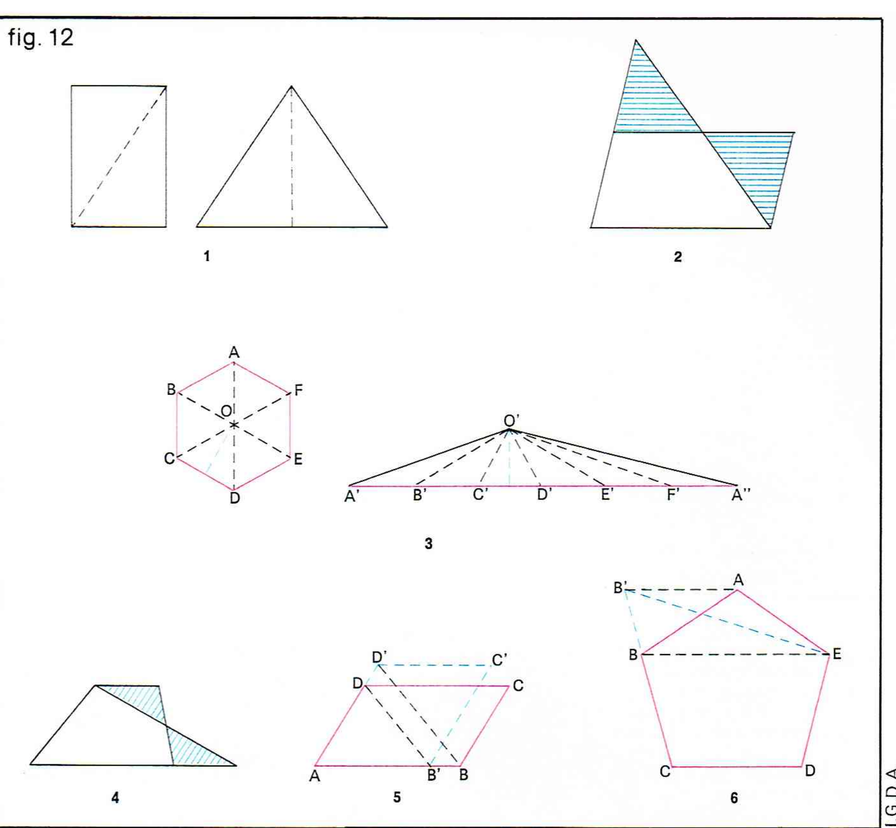
On démontre le théorème suivant très important : *deux figures symétriques sont isométriques.*

Grandeurs géométriques

Considérons un ensemble (généralement infini) d'êtres entre lesquels on peut établir une relation d'équivalence. Sur les classes d'équivalence, on se donne une opération interne qu'on appelle somme et une relation d'ordre total compatible. Il est donc permis de parler de multiple d'une grandeur. En pratique, chaque « grandeur » est à son tour une classe de figures géométriques équivalentes : les segments, les angles, les polygones sont des classes de grandeurs homogènes en identifiant l'équivalence avec la congruence dans les deux premiers cas, avec l'équidécomposabilité dans le troisième. On a l'habitude de parler de somme et de comparaison de figures géométriques, en sous-entendant de se référer aux grandeurs.

Lorsqu'on se fixe un élément arbitraire V dans une classe de grandeurs homogènes, le rapport *k* d'une grandeur G de la classe par rapport à V s'appelle mesure de G par rapport à l'unité V; et si on a une correspondance biunivoque entre *k* et G, on traduit toute relation d'équivalence, de somme par rapport à la mesure. Le choix de l'unité de mesure se fait en considération des besoins : l'unité de longueur est le *mètre*, celle de l'amplitude angulaire est le *degré* ou le *radian* (*fig.* 12, 13, 14).

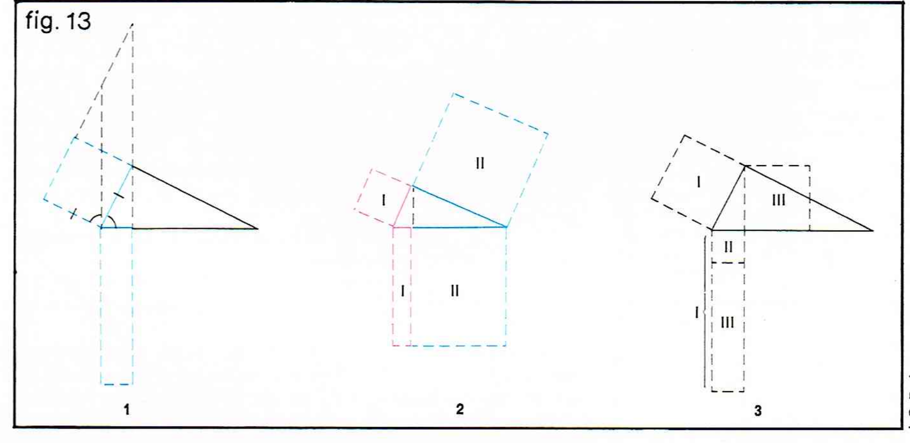
Par définition, deux classes K et K' de grandeurs respectivement homogènes, en correspondance biunivoque, sont dites proportionnelles si les rapports de deux gran-

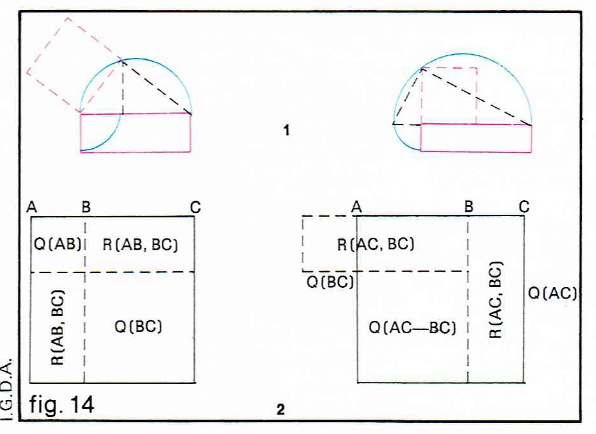


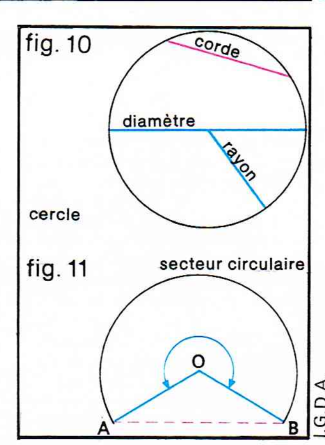
▲ **Figure 12** — ***polygones équivalents :*** **1)** ***équivalence de deux polygones décomposables en un même nombre de triangles respectivement égaux;*** **2)** ***équivalence d'un triangle avec un parallélogramme;*** **3)** ***équivalence d'un polygone régulier avec un triangle;*** **4)** ***équivalence d'un trapèze avec un triangle;*** **5)** ***équivalence de deux parallélogrammes;*** **6)** ***équivalence d'un polygone avec un autre polygone ayant un côté en moins.***

▼ ***Ci-dessous,*** **figure 13 — 1) premier théorème d'Euclide :** ***le carré qui a comme côté un côté de l'angle droit d'un triangle rectangle est équivalent au rectangle dont les côtés sont l'hypoténuse et la projection de ce côté sur l'hypoténuse;*** **2) théorème de Pythagore :** ***le carré construit sur l'hypoténuse d'un triangle rectangle est équivalent à la somme des carrés construits sur les côtés de l'angle droit;*** **3) deuxième théorème d'Euclide :** ***le carré construit sur la hauteur relative à l'hypoténuse d'un triangle rectangle est équivalent au rectangle qui a pour côtés les projections sur l'hypoténuse des deux côtés de l'angle droit.***

En bas, **figure 14 : 1)** ***les deux théorèmes d'Euclide permettent la construction du carré équivalent à un rectangle donné;*** **2)** ***le carré construit sur la somme (différence) de deux segments est équivalent à la somme des carrés construits sur chacun d'eux, augmentée (diminuée) du double du rectangle qui les a comme côtés.***

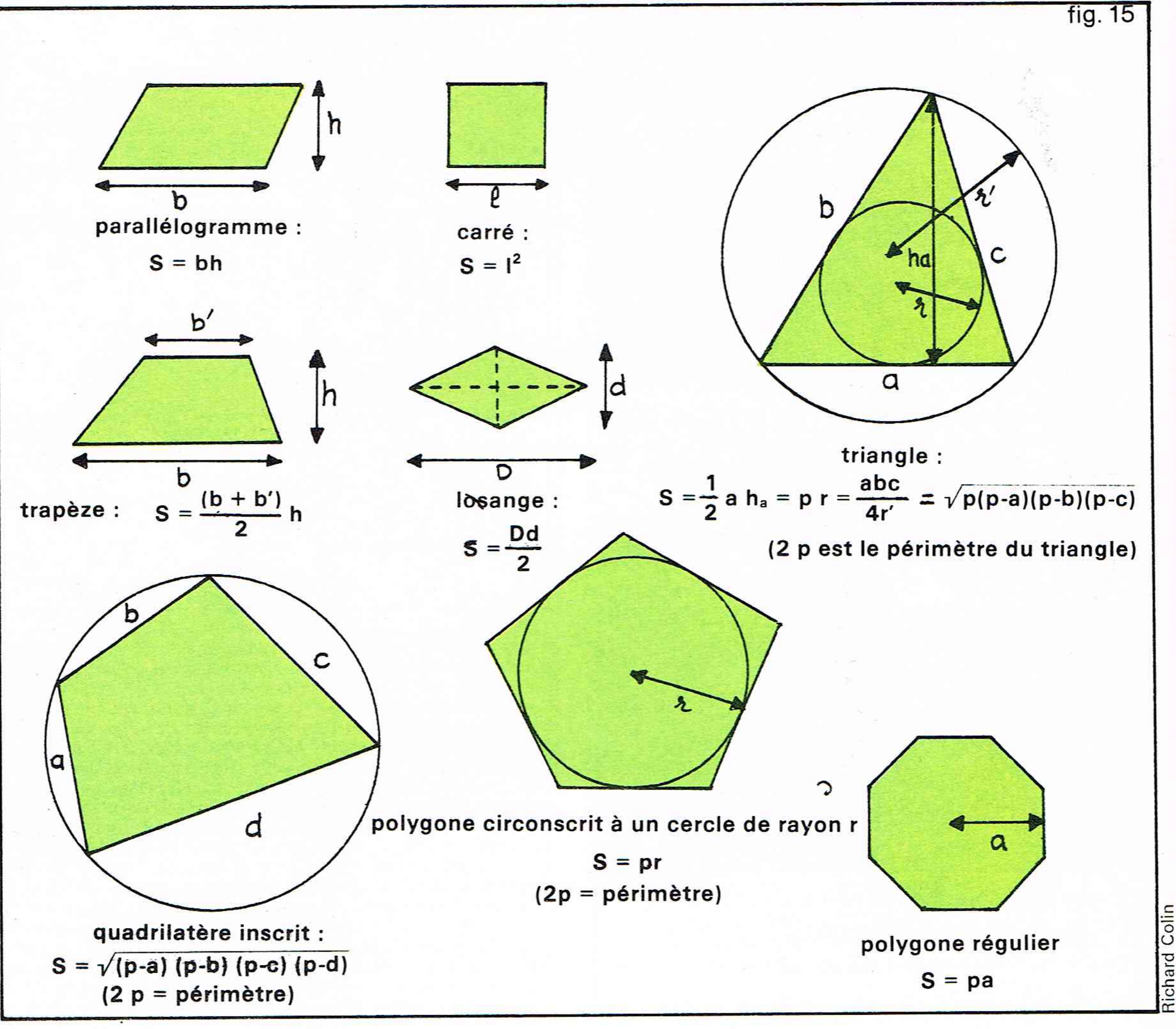






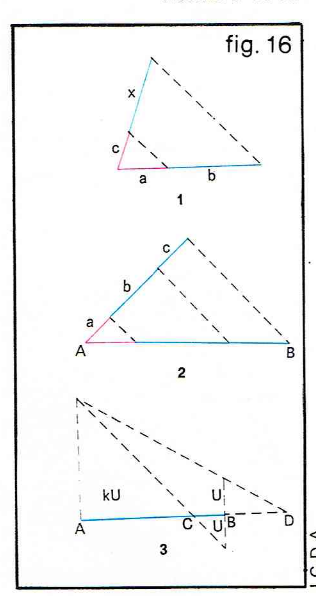


▲ *Le mathématicien et philosophe grec,* Thalès; *on lui doit avant tout d'avoir rapporté d'Égypte en Grèce les fondements de la véritable géométrie.*



▶ **Figure 15 :** *les mesures de surfaces des principales figures géométriques.*

▼ **Figure 16** — *conséquences du théorème de Thalès :* 1) *construction du quatrième segment proportionnel;* 2) *division d'un segment en parties proportionnelles à des segments donnés;* 3) *division d'un segment intérieurement et extérieurement dans le rapport k, où k est un nombre réel.*



deurs A, B de K et des correspondants A′, B′ de K′ sont toujours égaux.

L'égalité A : B = A′ : B′ s'appelle proportion et jouit des propriétés arithmétiques des proportions numériques.

On démontre le critère de proportionnalité suivant : une condition nécessaire et suffisante pour que deux classes K et K′ de grandeurs respectivement homogènes et en correspondance biunivoque soient proportionnelles est que :

— l'équivalence de deux grandeurs de K implique toujours celle des correspondants de K′;

— si, pour trois grandeurs A, B, C de K, on a : A = B + C, alors, pour les correspondants de K′, on a : A′ = B′ + C′.

En appliquant ce critère, on démontre que tous les arcs d'un cercle (ou tous les secteurs d'un cercle) et les angles correspondants au centre sont proportionnels.

De ces résultats, on peut déduire les mesures des surfaces des principales figures géométriques (*fig.* 15).

Les grandeurs homogènes dont nous venons de parler sont dites du *premier genre,* quand on identifie l'équivalence avec la congruence (c'est le cas des segments et des angles), et du *second genre* quand on identifie l'équivalence avec l'équidécomposabilité. On appelle grandeurs du *troisième genre* celles pour lesquelles l'équivalence n'est identifiable ni avec la congruence, ni avec l'équidécomposabilité. La méthode qui remonte à Eudoxe et qui fut utilisée systématiquement par Archimède pour traiter ces grandeurs est fondée sur le procédé d'« exhaustion » qui repose sur le postulat de la continuité.

Ce procédé a été utilisé pour mesurer la longueur d'un arc de cercle (et du cercle) et la surface d'un cercle. Il a été ainsi établi que le rapport entre un cercle et son diamètre est le nombre π, dont la transcendance a été démontrée par F. von Lindemann en 1882 :

π = 3,141 592 653 589 793 238 462 643 383 279 50...

(Archimède avait déjà réussi à démontrer que le nombre π était compris entre $3 + \frac{10}{71}$ et $3 + \frac{1}{7}$). On en déduit que la longueur d'un cercle de rayon r est égale à $2\pi r$. Le même procédé permet d'établir que la surface d'un cercle de rayon r est égale à πr^2.

La similitude

Du critère de proportionnalité, on déduit le *théorème de Thalès :* Un faisceau de droites parallèles détermine sur deux droites sécantes deux classes de segments proportionnels.

De ce théorème, on déduit que, si on se donne deux points A et B et un nombre réel k, il existe un seul point C intérieur au segment AB et, si $k \neq 1$, un seul point D sur son prolongement, tels que $AC/BC = k$, $AD/BD = k$ (*fig.* 16).

Il est usuel de dire que C et D divisent le segment AB intérieurement et extérieurement dans le rapport k; le quaterne ordonné ABCD est un groupe harmonique (voir *Géométrie projective*).

Deux figures sont dites semblables s'il est possible de trouver une correspondance biunivoque entre leurs points telle que deux segments homologues soient toujours proportionnels avec le même rapport k, appelé rapport de *similitude* (dans le cas particulier où ce rapport vaut 1, on a l'égalité).

Dans l'ensemble des figures d'un plan, la similitude est une relation d'équivalence.

Comme pour l'isométrie, on peut donner des définitions plus précises pour les polygones. Plus précisément : deux polygones sont semblables s'il existe entre leurs sommets une correspondance biunivoque telle que les angles homologues soient égaux et les côtés homologues proportionnels.

Pour les triangles, on a trois critères particuliers de similitude.

D. Ribas

▲ *Le cube qui a comme arête l'unité de mesure linéaire est pris comme unité de mesure du volume d'un polyèdre.*

La géométrie dans l'espace

Les axiomes

Tous les axiomes de la géométrie plane et leurs conséquences sont valables en géométrie dans l'espace. Il existe des axiomes supplémentaires *d'appartenance à l'espace* tels que : « étant donné un plan, il existe au moins un point qui ne lui appartienne pas et par conséquent une infinité » (*fig.* 17, 18, 19, 20, 21).

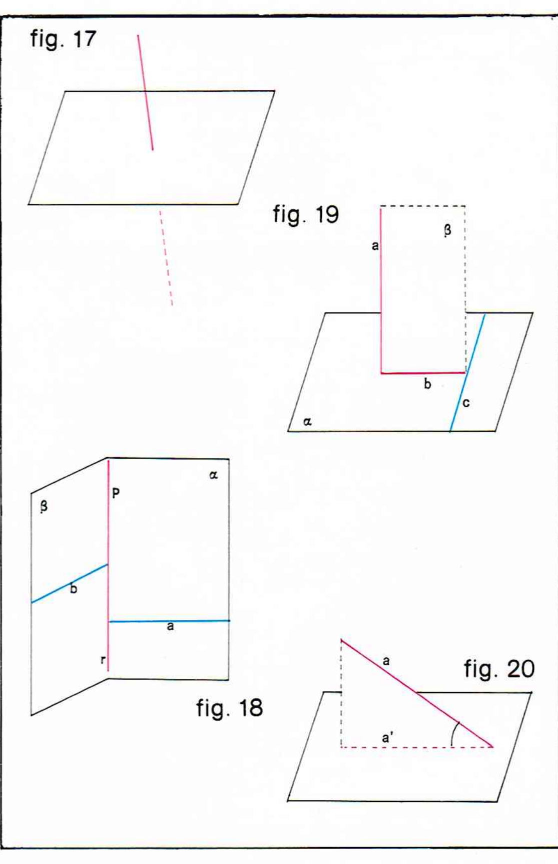
Tous les concepts de la géométrie plane tels que :
- l'orthogonalité
- le parallélisme
- la distance
- le lieu
- la symétrie

se généralisent en géométrie dans l'espace. Par exemple, on dit que : « deux plans sont parallèles s'ils n'ont pas de points communs » (*fig.* 21).

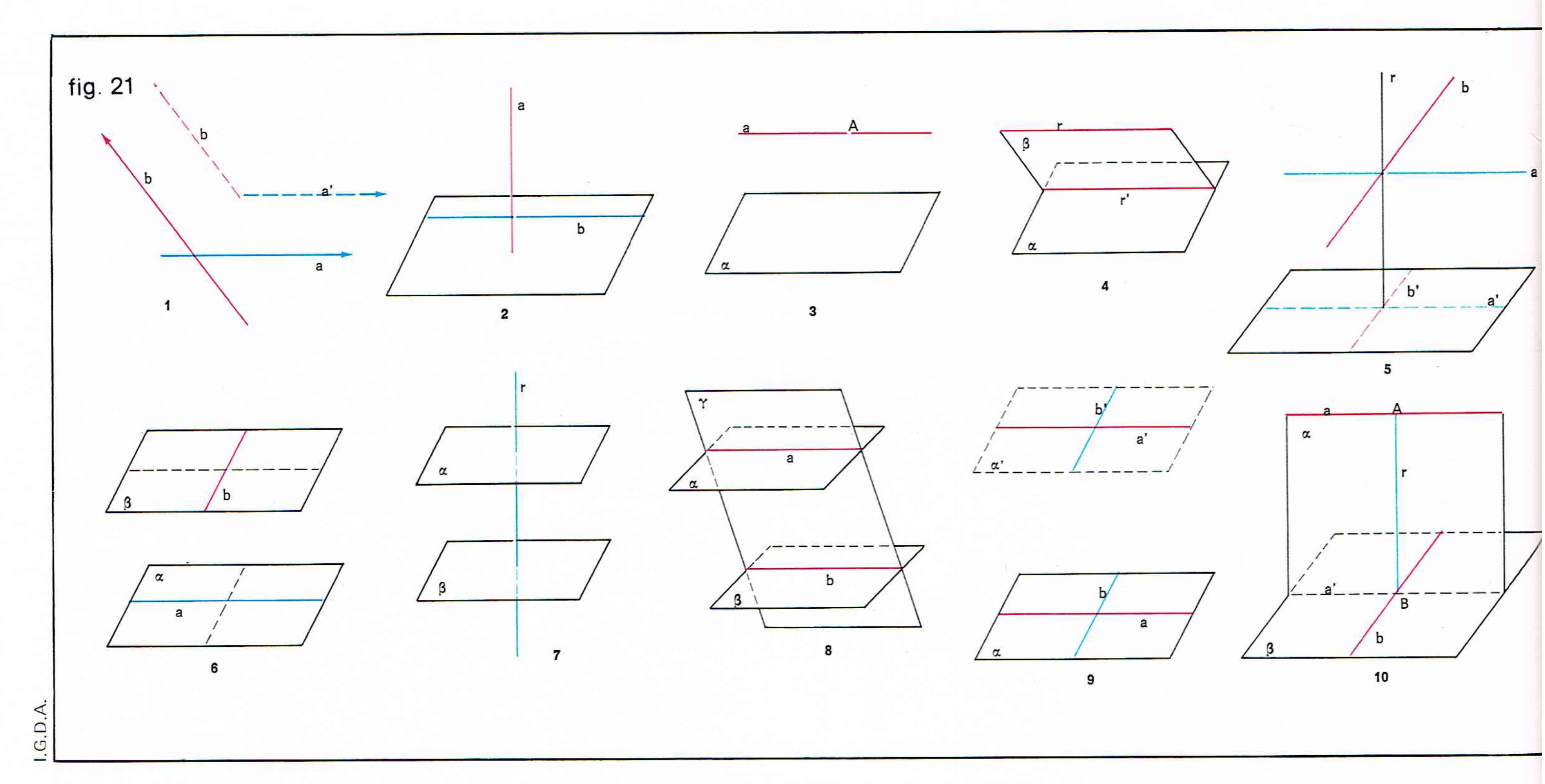
Principales figures de l'espace

Par définition, deux demi-plans ayant la même origine s déterminent deux régions, chacune d'elles s'appelant *dièdre* d'arête s et de faces α et β. Les concepts d'égalité, d'amplitude, d'inégalité, de somme de dièdres s'établissent de la même manière que pour les angles (*fig.* 22).

Considérons un polygone (plan) convexe à n côtés, un point V n'appartenant pas à son plan, et les n demi-droites ayant leur origine en V et passant chacune par un sommet du polygone ; l'ordre des sommets du polygone induit un ordre pour les demi-droites, de sorte qu'on peut prendre en considération les angles dont les côtés sont les couples de demi-droites consécutives. La surface constituée par l'ensemble de ces angles divise l'espace en deux régions ; celle qui ne contient pas les prolongements des demi-droites s'appelle *angle solide* convexe : V en est le sommet, les n demi-droites en sont les arêtes, tandis que les faces de l'angle solide sont les angles qui viennent d'être définis (*fig.* 23).



◄ ***Figure** 17 : demi-droite appartenant à un demi-espace déterminé par un plan.*
***Figure** 18 : par un point extérieur à deux droites non sécantes, passe une seule droite qui les intersecte toutes les deux.*
***Figure** 19 : théorème des trois perpendiculaires.*
***Figure** 20 : angle d'une droite avec un plan qu'elle intersecte.*



▲ **Droites et plans** *(figure 21)* **: 1,** ***angle de deux droites orientées; 2, droites orthogonales; 3, plan et droite parallèles; 4, plan passant par une droite parallèle à un plan; 5, plan et droites perpendiculaires à une droite en des points distincts; 6, unicité du plan passant par une droite et parallèle à une autre droite qui ne coupe pas la première; 7, plans perpendiculaires à une droite en des points distincts; 8, intersections parallèles d'un plan avec des plans parallèles; 9, plan parallèle à un autre, déterminé par 2 droites sécantes parallèles à 2 droites du plan donné; 10, distance de 2 droites non sécantes.***

fig. 22

I.G.D.A.

▶ **Dièdres** *(figure 22).* ***Plans perpendiculaires : 1, section normale d'un dièdre; 2, dièdre droit; 3, section inclinée d'un dièdre; 4, 5, 6, plans perpendiculaires; 7, unicité du plan perpendiculaire à un plan donné passant par une droite non perpendiculaire à ce dernier.***

On définit aussi les *polyèdres* qui sont des grandeurs du *troisième genre,* c'est-à-dire qu'il est possible de définir une relation d'équivalence dans l'ensemble des polyèdres, mais deux polyèdres équivalents ne sont généralement pas décomposables en un nombre fini de parties égales : la comparaison entre pyramides demande déjà le procédé d' « exhaustion ». On appelle *volume* l'abstraction relative à toute une classe de polyèdres équivalents. Comme unité de mesure du volume d'un polyèdre, on prend le cube qui a comme arête l'unité de mesure linéaire (*fig.* 24).

La géométrie moderne

En géométrie élémentaire, on a étudié des critères d'équivalence entre figures : on a parlé de figures égales, de figures équivalentes, de figures semblables. Dans la première partie du XIX[e] siècle, se développèrent d'autres types de géométrie (la plus importante étant la *géométrie projective*) : en résumé, elles établissaient de nouveaux critères d'équivalence entre figures. Le mathématicien *F. Klein* fut ainsi amené à exposer à l'université d'Erlangen un grand projet de classification des géométries, auquel fut précisément donné le nom de *programme d'Erlangen.* Il donna pour longtemps les idées de base des recherches en géométrie et montra la possibilité de construire de nouvelles géométries, qui furent ensuite étudiées avec succès. Plus récemment, le développement des recherches complique un peu le cadre général de la géométrie, mais le programme d'Erlangen reste toujours une œuvre des plus valables pour une vision plus claire et synthétique de la géométrie.

La géométrie et la théorie des groupes

Le *programme d'Erlangen* se réfère essentiellement aux espaces $\mathbb{R}^n$, $\mathbb{C}^n$ et $P_n(\mathbb{R})$ (voir *Topologie*), mais on peut l'appliquer à d'autres espaces.

L'idée centrale de F. Klein est la suivante : tout type de géométrie est caractérisé par une relation d'équivalence à laquelle on peut associer un groupe G de bijections de l'espace E sur lui-même (si E est un espace topologique, on parlera d'un groupe d'homéomorphismes [voir *Topologie*]).

On vérifie facilement que dans l'ensemble P (E) des parties de E, la relation « $A \sim B$ s'il existe une bijection f du groupe G telle que $B = f(A)$, $(A, B \in P(S))$ » est une relation d'équivalence :

si i est l'application identique telle que $i(x) = x$, on a $i(X) = X$, donc $X \sim X$;

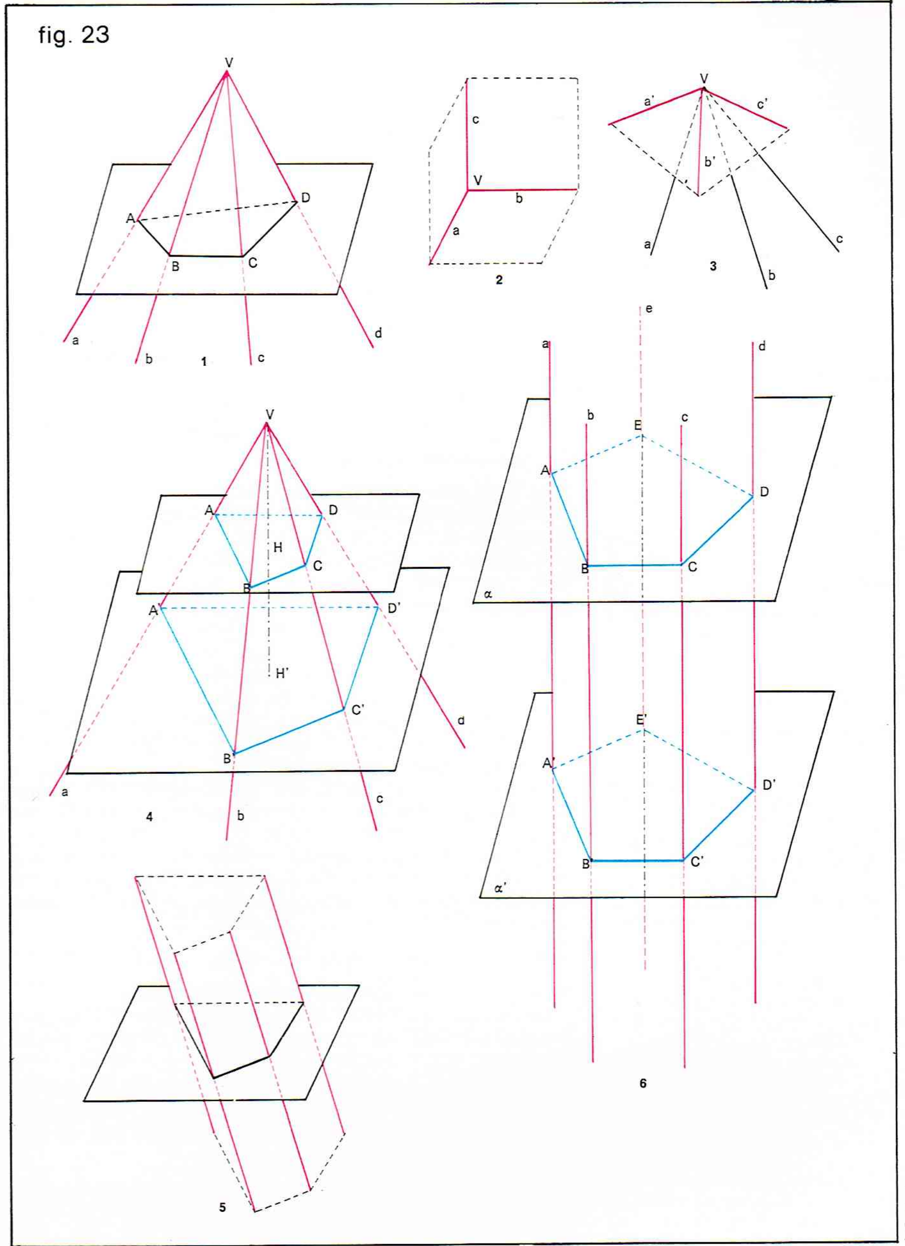
si Y = f (X) et Z = g (Y), on a Z = g (f (X)), et l'application composée de f et g étant encore biunivoque, de X ~ Y et Y ~ Z, on tire X ~ Z.

Un groupe G dans S donne donc la possibilité de parler de *figures équivalentes* respectivement à G, et d'établir la « géométrie du groupe G dans S ». Si G′ est un sous-groupe de G et si deux figures X et Y sont équivalentes par rapport à G′, alors elles le sont par rapport à G.

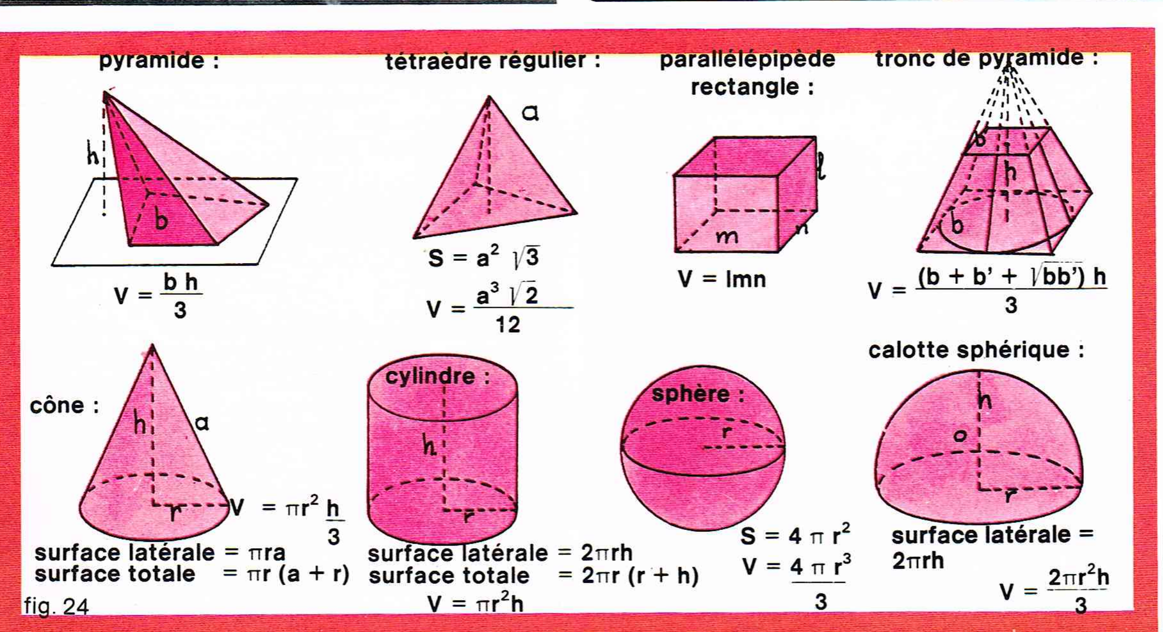
Dans la géométrie de G, cela a un sens de considérer une classe K d'objets si, par effet des correspondances de G, un objet de K se transforme en un objet de K : par exemple, on verra qu'en géométrie projective, on parle de l'ensemble des quadrangles (ou par abstraction, du concept de quadrangle : un quadrangle étant, par définition, une figure formée par quatre points et les six droites qui les joignent) puisque, par effet d'une *transformation projective*, un quadrangle devient un autre quadrangle ; mais on ne parle pas de parallélogramme puisqu'un parallélogramme peut se transformer en une autre figure. Tous les êtres qu'on peut considérer dans la géométrie de G peuvent aussi être considérés dans la géométrie de G′ si G′ ⊂ G : les correspondants de G′ sont en fait des correspondants particuliers de G. Par exemple, la notion de « triangle rectangle » appartient à la géométrie des similitudes et par conséquent à la géométrie métrique euclidienne ; inversement, la notion de périmètre est valable en géométrie métrique euclidienne, mais non en géométrie des similitudes (transformons par une similitude un polygone, on obtient un polygone de périmètre différent).

En général, on dit qu'une géométrie est *subordonnée* à une autre si le groupe de la première est un sous-groupe de celui de la seconde : naturellement, l'espace doit être


Palais de la Découverte, Paris



▲ *A gauche, le mathématicien* Félix Klein. *A droite, figure* 23 ; angles solides : 1, *angle solide convexe ;* 2, *trièdre trirectangle ;* 3, *trièdre polaire ;* 4, *similitude des polygones obtenus en coupant un angle solide par des plans parallèles ;* 5, *prisme convexe indéfini ou angle solide impropre ;* 6, *égalité des sections droites d'un même prisme.*



◀ *Figure 24 : on appelle* volume *l'abstraction relative à toute une classe de polyèdres équivalents.*

le même, néanmoins il est encore possible de considérer une subordination dans le cas où l'espace où opère le premier est contenu dans celui du second (on verra plus tard la relation entre la géométrie affine et la géométrie projective).

Soit F une figure; on appelle *invariant* de F dans la géométrie du groupe G (ou invariant par rapport à G) un nombre x (F), associé à F, qui a la même valeur pour toutes les figures transformées de F dans G, c'est-à-dire que $x(F) = x(g(F))$, pour tout $g \in G$. On dit inversement qu'une autre figure K (F) associée à F est *covariante* avec F (ou aussi qu'elle est invariante) lorsque

$$g(K(F)) = K(g(F))$$

(sa transformée par g coïncide avec la figure associée à g (F). Par exemple, le périmètre d'un polygone est un de ses invariants par rapport au groupe des isométries; le centre d'un cercle est une figure covariante avec le cercle lui-même par rapport au même groupe.

La géométrie projective

La table des matières du programme de F. Klein débute par la géométrie projective. Quelques-uns de ses aspects étaient connus depuis longtemps (on pourrait remonter aux géomètres grecs, comme *Apollonios* et *Pappus*, et plus récemment à *Desargues* et *Pascal*), mais elle fut systématiquement étudiée à partir du XIX[e] siècle (*Poncelet, Chasles, Gergonne, von Staudt* et les autres).

D'un point de vue analytique, on peut introduire la **géométrie projective** à partir de l'espace projectif $P_n(\mathbb{R})$ (voir *Topologie*) : les points y sont représentés par des $(n+1)$-uplets de nombres réels, non tous nuls (appelés coordonnées projectives homogènes), avec la convention que les coordonnées $(x_0, x_1, ..., x_n)$ et $(\rho x_0, \rho x_1, ..., \rho x_n)$, avec $\rho \neq 0$, représentent le même point; pour simplifier le raisonnement, on remplacera progressivement n par 2 (ou 3).

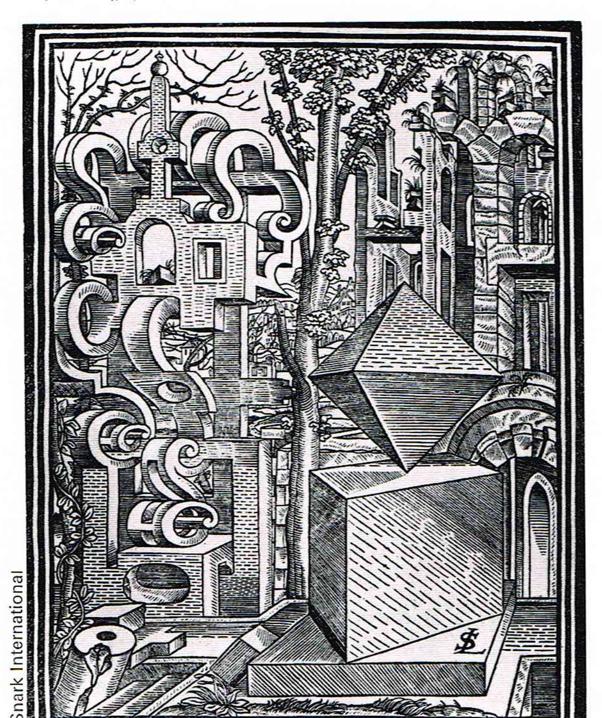
Comme ensembles particuliers de points de l'espace projectif, mentionnons les *hyperplans*, lieux des points dont les coordonnées satisfont à une équation du 1[er] degré homogène :

$$\textbf{(1)} \qquad u_0 x_0 + u_1 x_1 + ... + u_n x_n = 0$$

où les u_i sont des nombres réels non tous nuls.

Les u_i sont les coordonnées de l'hyperplan **(1)**, et deux hyperplans dont les coordonnées sont proportionnelles coïncident.

La géométrie projective est la géométrie du groupe des *homographies*, c'est-à-dire de l'ensemble des applications de $P_n(\mathbb{R})$ dans lui-même qui au point (x_i) associe le point (y_i) tel que :



Snark International

▶ *Page ci-contre, en bas; figure* **31** : *un* **quaterne harmonique**; *figure* **31'** : *un* **ordre circulaire**; *figure* **32** : **postulat d'ordre.**

▶ *Gravure extraite de la* **Géométrie** *de Giovanni Pomodoro (1624).*

$$y_i = a_{i0}x_0 + a_{i1}x_1 + ... + a_{in}x_n$$
$$(i = 0, 1, ..., n)$$

avec det $|a_{ik}| \neq 0$ (voir *Algèbre linéaire*).

Signalons un important invariant projectif (c'est-à-dire une quantité invariante par homographie), le birapport (A, B, C, D) de quatre points A, B, C, D d'une droite qui vaut : $\frac{\overline{CA}}{\overline{CB}} : \frac{\overline{DA}}{\overline{DB}}$. Ce birapport est aussi appelé rapport anharmonique et, dans le cas où il a pour valeur — 1, rapport harmonique.

On peut remplacer en géométrie projective le corps des nombres réels par le corps des nombres complexes ou un corps quelconque.

La géométrie projective peut aussi s'étudier par voie axiomatique, sans l'usage d'une méthode analytique.

En géométrie projective, on considère un espace comprenant une infinité d'objets appelés points, droites, plans liés par trois groupes de postulats :
- — les postulats d'appartenance;
- — les postulats d'ordre;
- — les postulats de continuité.

Observons avant d'énoncer ces postulats que, dans ce qui précède, les mots point et hyperplan peuvent être échangés. Cette remarque contient, en substance, le principe de *dualité* de *Poncelet* qu'on va appliquer en faisant correspondre à chaque proposition sa duale.

Les postulats d'appartenance :

[P1] *Deux points distincts déterminent une droite unique qui les contient.*	**[P1']** *Deux plans distincts déterminent une droite qui leur appartient.*
[P2] *Un point et une droite ne passant pas par ce point déterminent un plan qui les contient.*	**[P2']** *Un plan et une droite non contenue dans ce plan déterminent un point qui leur appartient.*

[P3] *Si un point appartient à une droite et celle-ci à un plan, le point appartient au plan.*

Indiquons deux de leurs conséquences immédiates :

1) *Deux droites distinctes se coupant en un point forment un plan (appelé conjugué des deux droites).*	1') *Deux droites appartenant à un plan se coupent en un point (appelé intersection des deux droites).*
2) *Trois points n'appartenant pas à une même droite déterminent un plan auquel ils appartiennent.*	2') *Trois plans ne passant pas par une même droite déterminent un point qui leur appartient.*

On remarque donc que deux droites coplanaires « se rencontrent » toujours, de même que deux plans.

Observons en outre qu'on passe des propositions écrites à gauche à celles écrites à droite en changeant les mots « point » et « plan ». On dit que [P_1], [P_1'] sont duaux. Quand on a démontré un théorème en géométrie projective, on a automatiquement prouvé son dual. Le postulat [P_3] qui coïncide avec son dual est dit autodual.

Pour exclure des exemples d'espaces trop réduits, on ajoute l'axiome suivant :

[P4] *Il existe au moins cinq points tels que quatre quelconques d'entre eux ne soient pas coplanaires.*

La géométrie projective considère six formes fondamentales :

1[re] espèce

Ponctuelle : ensemble des points d'une droite.	*Faisceau :* ensemble des plans passant par une droite *(axe)*.

Faisceau de droites : ensemble des droites d'un plan passant par un point *(centre)*.

2[e] espèce

Système plan : ensemble des points et des droites d'un plan.	*Gerbe :* ensemble des droites et des plans passant par un point.

3[e] espèce

Espace : considéré comme ensemble de ses points et de ses plans.

Mentionnons les applications fondamentales :

La projection des points d'une droite r du point O (n'appartenant pas à *r*) est la bijection, entre *r* et le faisceau de centre O situé dans le plan O*r*, qui à tout point P de *r* associe la droite OP (*fig.* 25).

La section d'un faisceau de plans avec un plan ω (n'appartenant pas au faisceau) est l'application duale de la projection de *r* sur O (*fig.* 26).

La projection des points d'une droite s sur une droite s ne coupant pas r est la bijection entre *r* et le faisceau d'axe *s* qui au point P associe le plan *s*P (*fig.* 27).

La section d'un faisceau de plans par une droite s non coplanaire avec l'axe du faisceau est l'application duale de la projection de *r* sur *s* (*fig.* 28).

Dans le plan, on a un autre type de dualité, la dualité plane : dans toute proposition de la géométrie plane, on peut remplacer les mots point et droite. Si F est une figure plane de points et droites, on la projette du point extérieur O (on obtient ainsi une figure de droites et plans), et ensuite on applique la dualité spatiale : on obtient une figure de droites et points.

Par exemple, dans le plan, on a l'opération duale par dualité plane de la projection de *r* de O : la section d'un faisceau avec une droite.

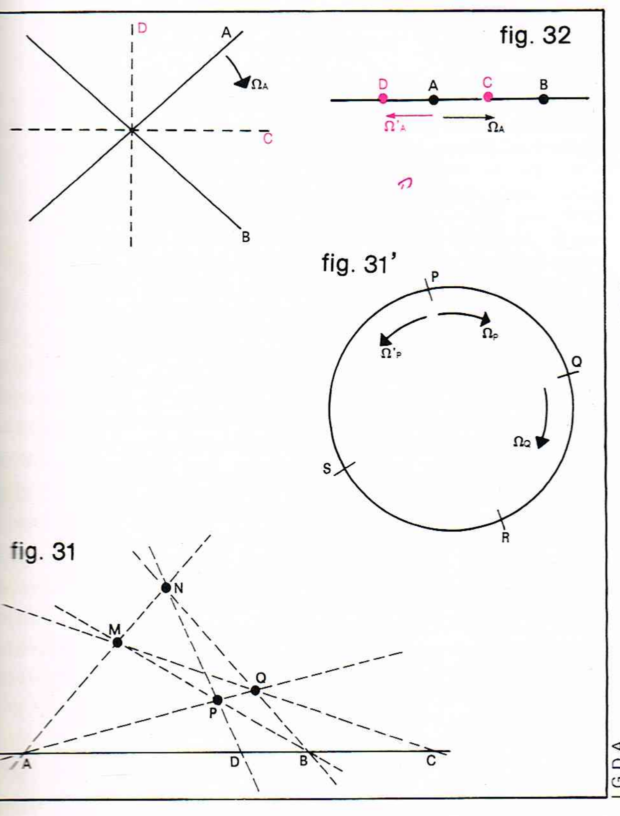
Le théorème des triangles homologiques est fondamental : si, dans un même plan ou dans l'espace, deux triangles ABC et A′B′C′ sont tels que les droites joignant respectivement les trois couples de sommets homologues AA′, BB′, CC′ se rencontrent en un même point S, les trois points de concours des couples de droites portant les côtés homologues des deux triangles sont alignés, et réciproquement (*fig.* 29).

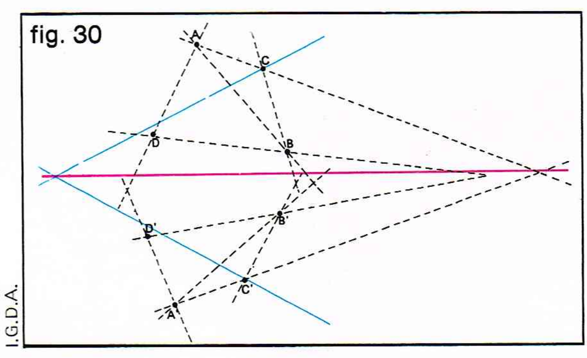
On a aussi un théorème des *quadrangles homologiques* (*fig.* 30).

Les points (A, B, C, D) d'une droite forment un *quaterne harmonique* s'il existe un quadrangle plan MNPQ de telle sorte que A soit l'intersection de MN et PQ, B celle de MP et NQ, et que C soit situé sur MQ et D sur NP : le théorème du quadrangle affirme que, les points A, B, C étant donnés, le point D est univoquement déterminé, indépendamment du choix du quadrangle MNPQ (*fig.* 31).

◄ *Figures* 25, 26, 27, 28 : *voir explications dans le texte.*

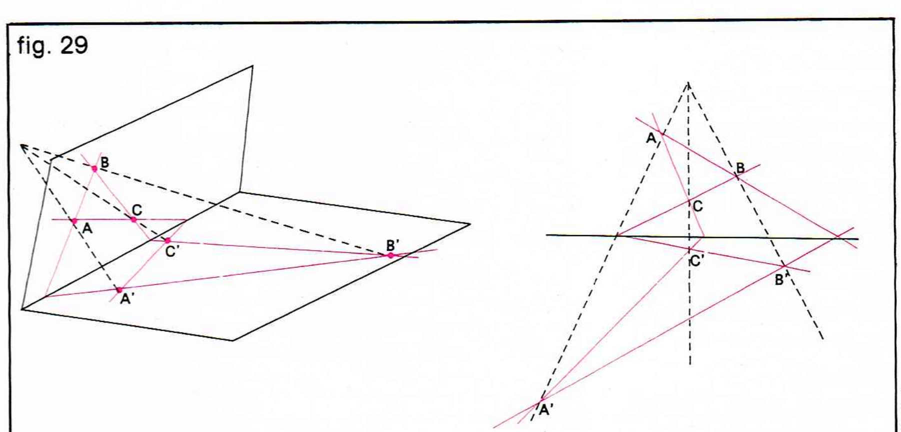
I.G.D.A.

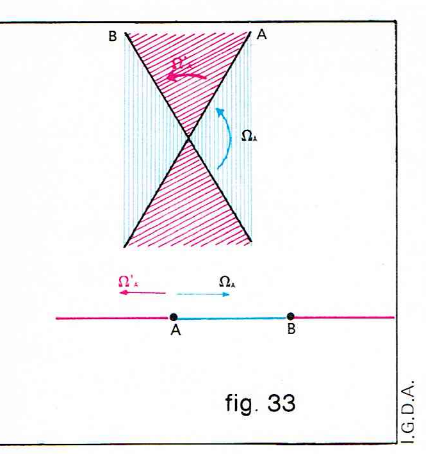




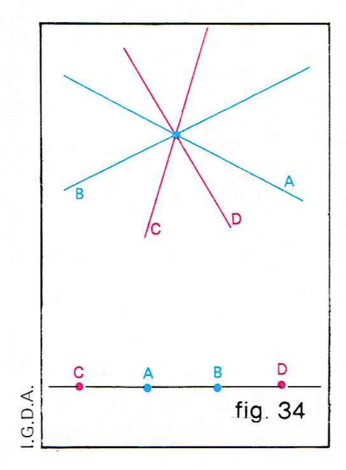
◄ *Figure* 30 : *théorème des* quadrangles homologiques.

▼ *Figure* 29 : *théorème des* triangles homologiques.





▲ ***Figure 33** : voir explications dans le texte.*



▲ ***Figure 34** : voir explications dans le texte.*

Les axiomes jusqu'ici introduits sont satisfaits aussi par des modèles ayant un nombre fini de points. Pour avoir une géométrie plus proche de l'intuition, on a besoin d'en ajouter d'autres.

Éclaircissons avant tout ce qu'on entend par *ordre circulaire* : on essaye en substance de caractériser la manière avec laquelle on suit les points d'un cercle parcouru dans un certain sens. Si on lui enlève un point P, il devient un segment. On peut ainsi dire qu'un ordre circulaire Ω dans un ensemble I est une famille Ω_p d'ordres totaux (un pour chaque $P \in I$) dans $I - \{P\}$, de telle manière que si dans Ω_p $Q < R < S$, alors dans Ω_Q : $R < S < P$ (*fig.* 31').

Nous appellerons *inverse* de Ω (et on désigne par Ω') l'ordre circulaire formé des ordres inverses des Ω_p.

Donnons alors un nouveau postulat (d'ordre) :

[P5] *Sur une forme de première espèce, il existe deux ordres circulaires, inverses l'un de l'autre, tels que, deux éléments A, B étant donnés, il existe au moins un élément C qui précède B dans Ω_A (fig. 32).*

Il s'ensuit qu'il existe aussi un élément D qui précède A dans Ω_B ; que Ω_A n'admet ni de minimum, ni de maximum, et que toute forme de première espèce (et aussi toute forme de 2e et de 3e espèce) a des éléments infinis.

Étant donné deux éléments A, B et l'ordre Ω_A, l'ensemble des éléments qui dans Ω_A précède B s'appelle un *segment* d'extrémités A, B. Si on considère aussi l'ordre Ω', on a deux segments d'extrémités A, B (ils sont dessinés avec des couleurs différentes sur la figure 33).

On admettra alors l'axiome (d'invariance des ordres) :

[P6] *L'image d'un segment par une projection et une section est un segment (c'est-à-dire que les ordres se conservent par projections et sections).*

Étant donné deux couples [A, B], [C, D] d'éléments distincts d'une forme de première espèce, il peut arriver que C, D (*fig.* 34) soient situés sur le même segment d'extrémités A, B (dans un tel cas, A et B aussi sont situés sur le même segment CD) [*fig.* 32]. Dans ce cas, on dit que les couples [A, B], [C, D] ne sont pas séparés, et dans le cas contraire qu'ils sont séparés (ce qui arrive, par exemple, lorsque A, B, C, D sont des éléments d'un quaterne harmonique).

Donnons enfin l'axiome de la continuité sous la forme de Dedekind :

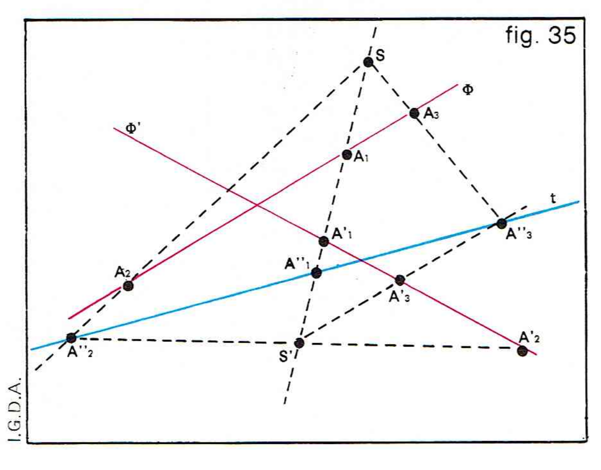
[P7] *Étant donné une partition d'un segment d'une forme de première espèce en deux classes, telles que (dans un des ordres de la forme) tout élément de la première classe précède tout élément de la seconde, alors il existe un élément maximal dans la première classe ou un élément minimal dans la seconde.*

Une projectivité entre des formes de 1re espèce est une bijection qui s'obtient en appliquant successivement un nombre fini de projections et de sections. Une projectivité entre deux formes de 2e espèce, deux systèmes plans, par exemple, fait correspondre à un point d'un des plans un point ou une droite de l'autre, ou aux points d'une droite du premier les points d'une droite ou les droites d'un faisceau de l'autre.

Dans le premier cas, la projectivité est une *homographie* ou *collinéation,* dans le second une *réciprocité* ou *corrélation*. Deux formes de 3e espèce sont projectives lorsque à un point de l'une correspond un point ou un plan de l'autre et qu'aux points d'une ponctuelle correspondent les points d'une ponctuelle ou les plans d'un faisceau... Dans le premier cas, on a une *homographie,* dans le second une *réciprocité*.

Les projectivités d'une forme assignée dans elle-même forment un groupe.

Sur une forme Φ d'espèce k ($k = 1, 2, 3$), déterminons $k + 2$ éléments distincts A_i (pour $k = 2, 3$, supposons que 3 d'entre eux n'appartiennent pas à une forme de 1re espèce) ; sur une autre forme Φ' de la même espèce, déterminons $k + 2$ éléments A_i, avec les mêmes conditions. Il existe une et une seule projectivité entre les deux formes qui à A_i associe A_i'. Dans l'exemple de la figure 35, on a reporté la construction d'une telle projectivité dans le cas où Φ et Φ' sont des ponctuelles (la séquence d'opérations, à partir de Φ, est la suivante : projections de S, section avec t, projections de S', sections avec Φ'). La construction est une preuve directe de l'existence d'une telle projectivité.



▶ ***Figure 35** : projectivité entre deux ponctuelles (Φ et Φ').*

La géométrie projective peut aussi se construire à partir de la géométrie euclidienne (ou plus précisément affine). Considérons les directions de l'espace ordinaire (c'est-à-dire les ensembles de droites deux à deux parallèles), et appelons-les *points impropres,* tandis que les points de l'espace sont appelés *points propres*.

On appelle ensuite *droite impropre* l'ensemble des points impropres d'un plan, et *plan impropre* l'ensemble de tous les points impropres. On constate facilement que les axiomes de la géométrie projective sont vérifiés. Toute droite propre acquiert un point ultérieur, son point impropre, qu'on « rejoint » en parcourant la droite dans l'un ou l'autre de ses sens; la droite devient une *courbe fermée,* comme un faisceau de droites.

La géométrie affine et ses subordonnées

La **géométrie affine,** tout en n'ayant pas eu l'importance historique de la géométrie projective, est assez importante par de nombreux aspects. Pour donner une idée générale, on peut dire qu'en géométrie affine, outre toutes les notions d'appartenance, on garde celle de parallélisme, mais non celle de perpendiculaire, distance, etc.

D'un point de vue analytique, l'espace affine est l'espace $\mathbb{R}^n$, c'est-à-dire l'ensemble des n-uplets de nombres réels, muni du groupe des affinités, qui au point x fait correspondre le point y tel que :

$y_i = a_{i1} x_1 + a_{i2} x_2 + \ldots + a_{in} x_n + b_i, \quad (i = 1, 2, \ldots, n)$

avec $\det |a_{ik}| \neq 0$.

En géométrie affine, deux parallélogrammes quelconques sont équivalents, tandis que ne le sont pas un parallélogramme et un quadrilatère qui n'est pas un parallélogramme. Parmi les coniques, deux ellipses sont équivalentes, de même que deux hyperboles et deux paraboles; mais une ellipse et une hyperbole ne sont pas équivalentes. On ne parle pas de rectangle, de losange, de cercle, puisque dans la définition de telles figures interviennent des concepts étrangers à la géométrie affine.

On peut aussi parler de géométrie affine sur un corps, en considérant comme espace l'ensemble des n-uplets ordonnés d'éléments du corps. On peut aussi passer de la géométrie affine à la géométrie projective. Dans l'espace affine, introduisons les *points impropres,* on obtient un espace projectif. Réciproquement, étant donné un espace projectif, on fixe un de ses hyperplans α comme *hyperplan impropre :* appelons parallèles les hyperplans qui se coupent en α et appelons affinités les homographies qui conservent α : on a alors un *espace affine*.

La géométrie affine peut aussi se définir par voie axiomatique; il suffit de modifier à bon escient quelques axiomes de la géométrie projective (en particulier, au lieu de **[P5]**, on postulera l'existence sur une droite de deux ordres totaux, l'un inverse de l'autre).

On peut définir, à partir d'un espace affine, un espace vectoriel de la manière suivante. Considérons les couples ordonnés de points (*fig.* 36), et établissons entre eux la relation d'équivalence suivante *(équipollence) :* si A, B, X, Y sont des points de coordonnées a_i, b_i, x_i, y_i *(i variant de 1 à n),* on pose :

$$(A, B) \sim (X, Y) \text{ si } b_i - a_i = y_i - x_i \text{ quel que soit } i.$$

Les classes d'équivalence s'appellent les *vecteurs libres* de l'espace. Les différences indiquées s'appellent provisoirement *composantes* du vecteur libre. Soit deux vecteurs u, v de composantes (u_i), (v_i) et un nombre réel k (si l'espace est défini sur un autre corps, un élément du corps) ; on définit le vecteur $u + v$ comme le vecteur de composantes $(u_i + v_i)$ et le vecteur ku comme le vecteur de composantes (ku_i). De cette manière, l'ensemble des vecteurs libres est muni d'une structure d'espace vectoriel sur $\mathbb{R}$ (ou plus généralement, sur le corps sur lequel est défini l'espace affine). Les (u_i) sont les composantes du vecteur u par rapport à la base formée des vecteurs de composantes (*fig.* 37) :

$$(1, 0, ..., 0), \quad (0, 1, ..., 0), \quad ..., \quad (0, 0, ..., 0, 1).$$

Pour d'autres applications, il peut être utile de considérer dans l'ensemble des couples de points d'autres relations d'équivalence : par exemple, celle qui associe des couples ordonnés équipollents et appartenant à la même droite.

Parmi les affinités d'un espace sur lui-même, notons en particulier :

— les *translations* qui forment un groupe, qu'on peut identifier au groupe additif des vecteurs libres ;

— *la symétrie par rapport à un point A*. Les symétries ne forment pas un groupe, mais l'ensemble des translations et des symétries forme un groupe ;

— *les homothéties :* on les obtient en « dilatant » ou en « contractant » l'espace par rapport à un de ses points appelé centre. Si les coordonnées du centre sont a_i, les homothéties sont données par :

$$y_i = px_i + b_i \quad (p \neq 0 \quad \text{et} \quad p \neq 1)$$

avec $b_i = a_i\,(1 - p)$. Les homothéties de centre donné (application identique comprise) forment un groupe, ainsi que toutes les homothéties et les translations de l'espace.

On a des géométries subordonnées à la géométrie affine : la **géométrie des similitudes** fondée sur le groupe des similitudes qui est un sous-groupe de celui des affinités, et la **géométrie métrique euclidienne** subordonnée à la géométrie semblable puisque son groupe de transformation, qui est celui des isométries, est un sous-groupe de celui des similitudes.

Fixons dans l'espace affine à n dimensions un n-uplets e_i de vecteurs linéairement indépendants, c'est-à-dire une base de l'espace vectoriel associé (voir *Algèbre linéaire*) ; si on se fixe une origine O, on a un système de référence : à tout point P, on associe comme coordonnées les composantes du vecteur d'extrémités O, P. Toute autre base e'_i qui s'obtient ainsi à partir de e_i : $e'_i = \sum_{k=1}^{n} a_{ik}\, e_k$ avec :

$$\textbf{(1)} \qquad \sum_{i=1}^{n} a^2_{ik} = 1 \qquad \sum_{i=1}^{n} a_{ik}\, a_{ik'} = 0 \text{ pour } k \neq k'$$

sera dite orthonormale. Les similitudes sont les bijections de $\mathbb{R}^n$ dans lui-même représentées par :

$$y_i = \rho \sum_{k=1}^{n} a_{ik}\, x_k + b_i,$$

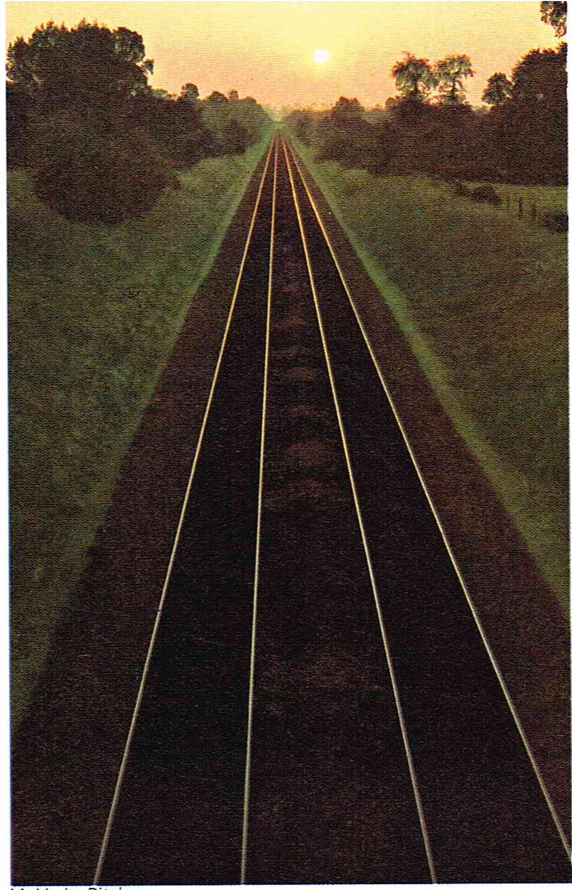
où les a_{ik} satisfont à **(1)**, et ρ est un nombre réel positif. Pour $n = 1$, on retrouve toutes les affinités de $\mathbb{R}$; pour $n = 2$, on a les correspondances qui à (x_1, x_2) associent (y_1, y_2) :

$$y_1 = \rho\,(x_1 \cos \alpha \mp x_2 \sin \alpha) + b_1$$
$$y_2 = \rho\,(x_1 \sin \alpha \pm x_2 \cos \alpha) + b_2 ;$$

les similitudes correspondant aux signes supérieurs sont dites directes, les autres inverses. Toutes les similitudes, ainsi que les seules similitudes directes, forment un groupe.

Les isométries sont les similitudes pour lesquelles on a : $\rho = 1$. Les isométries forment un groupe, ainsi que les Isométries directes : elles transforment une base orthonormale en une base orthonormale.

Un couple de droites distinctes non parallèles admet un invariant par rapport au groupe des similitudes : leur angle ; un couple de points distincts n'en admet pas. Par contre, un tel couple admet un invariant par rapport au groupe des isométries : leur distance. Pour le traité axiomatique de la géométrie euclidienne, voir la *Géométrie élémentaire*.



M. Varin-Pitch

Aperçu sur les géométries non euclidiennes

On appelle **géométrie non euclidienne** une géométrie de caractère métrique dans laquelle le *postulat des parallèles* ou *postulat d'Euclide* est remplacé par un axiome non équivalent. Ces géométries sont de deux types.

— *La géométrie hyperbolique,* dans laquelle on admet que, dans un plan donné, par un point extérieur à une droite, on peut mener une infinité de parallèles. Cette géométrie fut introduite par *N. Lobatchevski* au début du XIX[e] siècle.

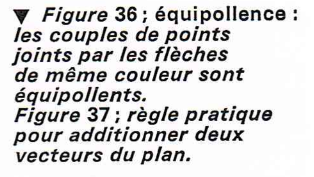
— *La géométrie elliptique,* dans laquelle deux droites coplanaires se coupent toujours. Pour pouvoir admettre un tel postulat, il est ainsi nécessaire de laisser tomber aussi l'axiome d'Archimède. Cette géométrie est due à *B. Riemann.*

De telles géométries peuvent se traiter de diverses façons. Ici, nous n'en montrerons qu'une pour la géométrie hyperbolique.

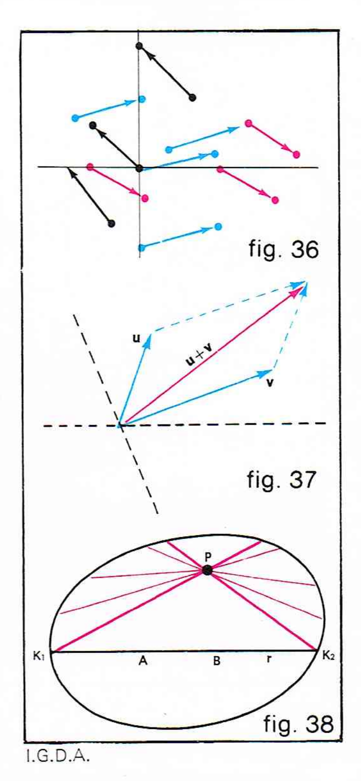
Dans le plan projectif réel, considérons une conique $\mathcal{C}$ de points réels et non dégénérée. On a un modèle de géométrie hyperbolique plane en prenant comme « points » les points intérieurs à $\mathcal{C}$ et comme « droites » les segments de droites intérieurs à $\mathcal{C}$. Il est immédiat de constater que, étant donné une « droite » r et un « point » P qui lui est extérieur, toutes les « droites » passant par P contenues dans un certain angle fermé n'intersectent pas r. Comme « distance » entre A et B, on peut prendre le nombre $k$lg (A, B, K_1, K_2), où k est une constante à fixer une fois pour toutes, et K_1, K_2 les intersections de la droite AB avec $\mathcal{C}$ (pour la notion de birapport, voir la *Géométrie projective*). Le groupe fondamental est celui des homographies qui transforment la conique $\mathcal{C}$ en elle-même (celles-ci conservent la distance définie plus haut) [voir *fig.* 38].

BIBLIOGRAPHIE

ARTIN E., *Algèbre géométrique,* trad. par M. Lazard, Gauthier-Villars, Paris, 1967. - CHOQUET G., *l'Enseignement de la géométrie,* Hermann, Paris, 1964. - DELACHET, *la Géométrie élémentaire,* P.U.F., Paris, 1966, collection « Que sais-je ? », n° 1211 ; *la Géométrie contemporaine,* P.U.F., Paris, 1969, collection « Que sais-je ? », n° 401. - DIEUDONNÉ J., *Algèbre linéaire et Géométrie élémentaire,* 3[e] édition, Hermann, Paris, 1968. - DONEDDU A., *Géométrie euclidienne plane,* Dunod, Paris, 1968. - HILBERT D., *les Fondements de la géométrie,* édition critique avec introduction et compléments préparée par P. Rossier, Dunod, Paris, 1971.

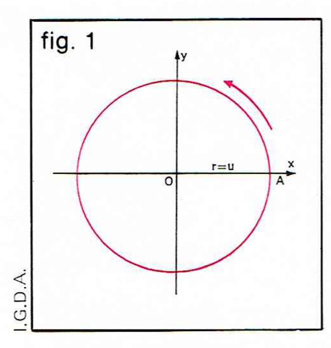


▼ ***Figure 36 ; équipollence : les couples de points joints par les flèches de même couleur sont équipollents.***
Figure 37 ; règle pratique pour additionner deux vecteurs du plan.



I.G.D.A.

▲ ***Figure 38 : modèle de géométrie hyperbolique plane obtenu en prenant comme « points » et « droites » les points et les segments de droites intérieurs à la conique.***



▲ *Figure 1 : le cercle trigonométrique.*

TRIGONOMÉTRIE

Le mot « trigonométrie » signifie étymologiquement « mesure des triangles ». En fait, l'objet de cette science est de faire le lien entre des grandeurs linéaires (longueurs, surfaces) et des grandeurs angulaires, par l'intermédiaire des *fonctions circulaires*. Elle doit son origine à des problèmes d'arpentage et d'astronomie.

On distingue la *trigonométrie plane,* qui ne traite en général que des figures planes, et la *trigonométrie sphérique,* plus compliquée, dont l'objet est l'évaluation des éléments de certaines figures de l'espace.

Les études théoriques commencées par les Babyloniens et les Grecs (Hipparque, Ptolémée), furent poursuivies d'abord par les Arabes, et ensuite en Europe par Regiomontanus, Copernic, Viète. Neper, avec la découverte des logarithmes (1614), donna à la trigonométrie de nouvelles possibilités, et de nouveaux développements furent trouvés par les mathématiciens des XVIIe et XVIIIe siècles : Legendre, Cauchy, J. Bernoulli, Newton, Leibniz, Delambre ; et le grand mérite d'avoir porté la théorie à sa version définitive appartient à Euler (1707-1783).

Théorie des fonctions trigonométriques

Étant donné un plan, choisissons un point quelconque O du plan comme origine et un segment u comme unité de mesure de longueur ; on choisit un système d'axes cartésiens orthogonaux $\overrightarrow{Ox}$ et $\overrightarrow{Oy}$ d'origine O (voir *Géométrie analytique*) avec des sens positifs tels que l'on passe de $\overrightarrow{Ox}$ à $\overrightarrow{Oy}$ par une rotation contraire au sens des aiguilles d'une montre, appelée sens trigonométrique (*fig.* 1).

Tout point du plan, et de là, tout point P d'un cercle de centre O et de rayon u, peut être repéré à l'aide de deux nombres qui s'obtiennent en mesurant avec u les deux segments x et y, projections de P sur les axes $\overrightarrow{Ox}$ et $\overrightarrow{Oy}$, et s'écrira : $P = P(x, y)$. Les angles ont comme origine le demi-axe positif $\overrightarrow{Ox}$ et comme sens positif le sens trigonométrique.

Comme unité de mesure angulaire, on utilise le *degré :* $1° = 1/360$ de l'angle égal à un tour complet, avec ses deux sous-multiples, la minute (1′) et la seconde (1″) (60′ = 1°, 60″ = 1′) ; le *radian :* 1 rd = angle qui représente sur le cercle un arc de longueur égale au rayon du cercle ; le degré est utilisé dans les applications pratiques, le radian dans les études théoriques.

La partie du plan limitée par le cercle de centre O et de rayon $r = u$ se nomme *cercle trigonométrique*.

Puisqu'on utilise fréquemment les deux unités angulaires, il est important de voir rapidement la relation qui les lie et les valeurs qu'ont les angles les plus importants mesurés avec ces unités. Pour avoir la mesure en radians d'un angle, il suffit de diviser un arc qu'il sous-tend par le rayon du cercle auquel appartient cet arc. Étant donné que la longueur d'un cercle de rayon r est $c = 2\pi r$, on a immédiatement :

un angle égal à un tour complet = 360° = 2 π rd.

Pour passer d'un système de mesure à l'autre, on utilise alors simplement la proportion :

$$\alpha°/\alpha \text{ rd} = 360°/2\pi$$

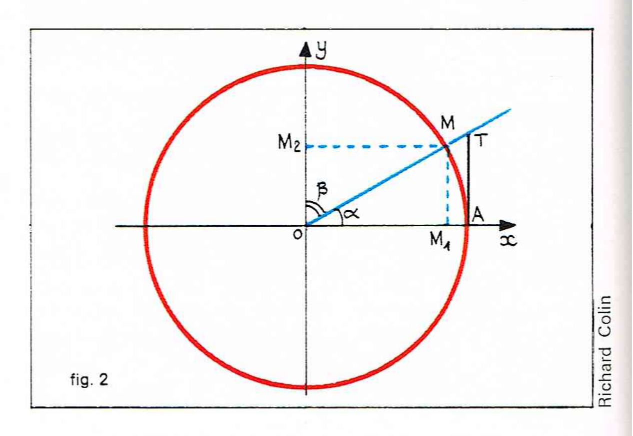
Tableau des unités de mesure de quelques angles remarquables

Degrés	**360**	**270**	**180**	**120**	**90**	**60**	**30**
Radians	**2π**	**3π/2**	**π**	**2π/3**	**π/2**	**π/3**	**π/6**

Les fonctions trigonométriques circulaires directes

On distingue *les fonctions circulaires directes* qui mesurent un segment en fonction d'un angle et *les fonctions circulaires inverses* qui mesurent un angle en fonction d'un segment. Sauf mention contraire, on mesurera dans cette partie les angles en radians.

Considérons (*fig.* 2) un cercle trigonométrique et désignons par α un angle orienté dont le premier côté coïncide avec le demi-axe $\overrightarrow{Ox}$; soit M l'intersection de l'autre côté de l'angle avec le cercle. Les fonctions qu'on va définir mettent en relation l'angle qui définit M avec trois segments. Désignons par A l'intersection du cercle avec $\overrightarrow{Ox}$, par M_1 et M_2 les projections de M sur les axes $\overrightarrow{Ox}$ et $\overrightarrow{Oy}$, et par T l'intersection de la tangente en A au cercle et de l'autre côté de l'angle ; on pose :



$$\sin\alpha = \frac{M_1M}{OA} \qquad \cos\alpha = \frac{OM_1}{OA} \qquad \text{tg}\,\alpha = \frac{AT}{OA}$$

c'est-à-dire qu'on définit :

le **sinus** d'un angle α (ou d'un arc de cercle $\widehat{AM}$) comme le nombre positif ou négatif qui mesure le segment M_1M par rapport au rayon ;

le **cosinus** d'un angle α (ou d'un arc $\widehat{AM}$) comme le nombre positif ou négatif qui mesure le segment OM_1 par rapport au rayon ;

la **tangente** d'un angle α (ou d'un arc $\widehat{AM}$) comme le nombre positif ou négatif qui mesure le segment AT par rapport au rayon.

Étant donné que les segments qui interviennent dans les définitions des fonctions trigonométriques sont parallèles à un des deux axes cartésiens, on peut leur attribuer un signe.

Comportement et graphes de ces fonctions : on peut tracer les graphes des trois fonctions $\sin x$, $\cos x$ et $\text{tg}\, x$. Sin x et cos x sont des fonctions définies sur $\mathbb{R}$ et périodiques de période 2π, c'est-à-dire que :

$$\sin(x + 2\pi) = \sin x \quad \text{et} \quad \cos(x + 2\pi) = \cos x ;$$

la tangente définie pour $x \neq (2k + 1)\frac{\pi}{2}$ (k entier), est périodique de période π :

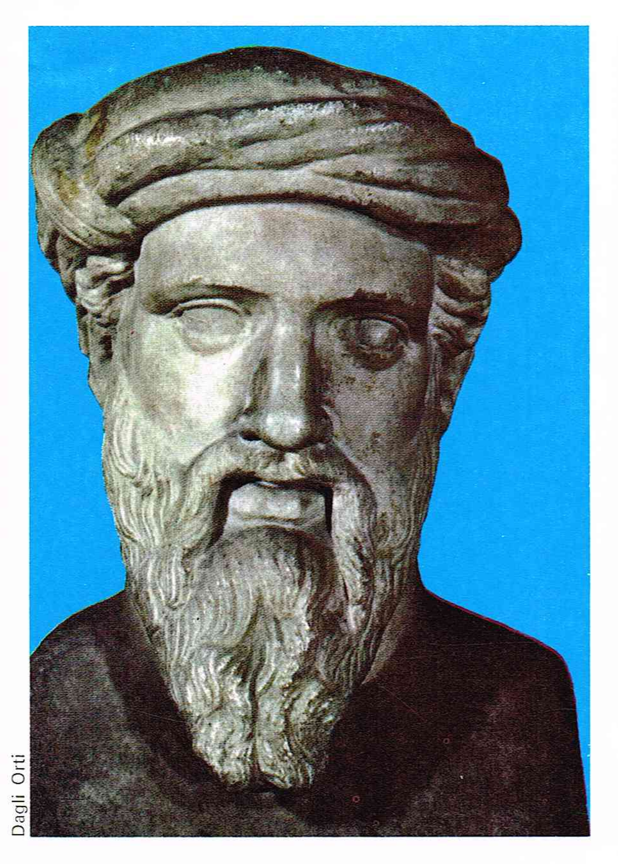
$$\text{tg}\,(x + \pi) = \text{tg}\, x.$$

Il est donc suffisant de les représenter sur un intervalle de longueur égale à leur période. Faisons décrire à M un tour complet du cercle trigonométrique à partir de A, on obtient alors les différentes valeurs des fonctions pour construire les trois graphes. Ces fonctions sont continûment dérivables (voir *Analyse*), ce qui facilite l'étude de leurs sens de variation (*fig.* 3, 4, 5) :

$$(\sin x)' = \cos x$$
$$(\cos x)' = -\sin x$$
$$(\text{tg}\, x)' = 1 + \text{tg}^2 x = \frac{1}{\cos^2 x}$$

On utilise aussi une quatrième fonction trigonométrique : la cotangente, définie ainsi :

$$\text{cotg}\, x = \text{tg}\left(\frac{\pi}{2} - x\right) ;$$



Dagli Orti

◀ *Le philosophe et mathématicien* Pythagore *(VIe siècle av. J.-C.).*

définie pour tout $x \neq k\pi$ ($k \in \mathbb{N}$), elle est périodique de période π comme la tangente, et sa dérivée est égale à :

$$(\text{cotg}\ x)' = -\frac{1}{\sin^2 x} = -(1 + \text{cotg}^2\ x)$$

En examinant le cercle trigonométrique, on peut trouver une relation qui ramène à deux le nombre des fonctions circulaires directes indépendantes ; en effet :

$$\cos \alpha = \frac{OM_1}{OA} = \frac{M_2M}{OA} = \sin \beta$$

(si on regarde la figure tournée de 90°), ce qui entraîne que :

$$\cos \alpha = \sin \left(\frac{\pi}{2} - \alpha\right).$$

Les coordonnées d'un point M d'une circonférence de rayon unité sont donc le cosinus et le sinus de l'angle qui sous-tend l'arc $\widehat{AM}$.

Les deux triangles OMM_1 et OTA étant semblables, on a :

$\dfrac{M_1M}{OM_1} = \dfrac{AT}{OA}$, ce qui entraîne que : $\text{tg}\ \alpha = \dfrac{\sin\ \alpha}{\cos \alpha}$.

En appliquant le théorème de Pythagore au triangle OMM_1, on obtient :

$$\sin^2 \alpha + \cos^2 \alpha = 1\ ;$$

et en divisant les deux membres par $\cos^2 \alpha$:

$$1 + \text{tg}^2\ \alpha = \frac{1}{\cos^2 \alpha}$$

Ce sont les trois relations fondamentales auxquelles on a souvent recours dans la théorie et les applications. On déduit facilement de ces relations :

$$\text{cotg}\ \alpha = \frac{\cos \alpha}{\sin \alpha} \qquad \text{et} \qquad 1 + \text{cotg}^2\ \alpha = \frac{1}{\sin^2 \alpha}.$$

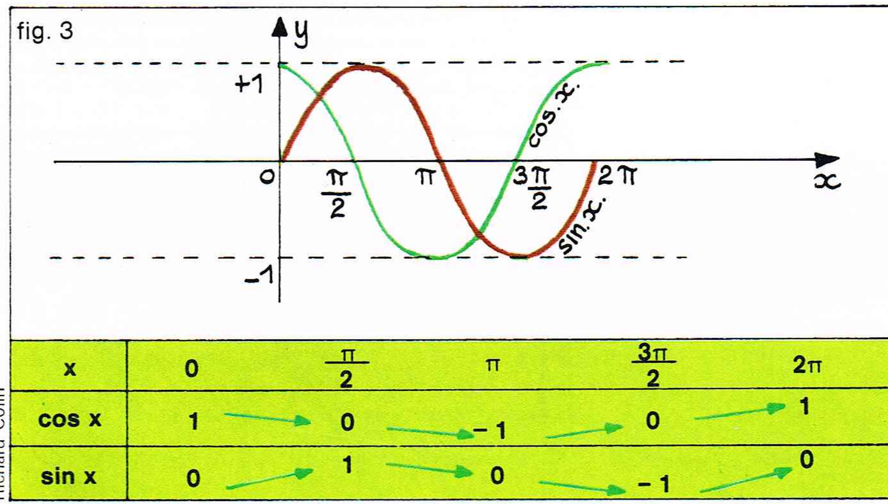
Réduction au premier quadrant

On appelle **réduction au 1er quadrant** l'opération qui consiste à ramener toute fonction trigonométrique d'un angle quelconque à une fonction trigonométrique d'un angle $< \frac{\pi}{2}$. En tenant compte de la propriété de périodicité

Tableau des valeurs des fonctions trigonométriques pour quelques angles remarquables

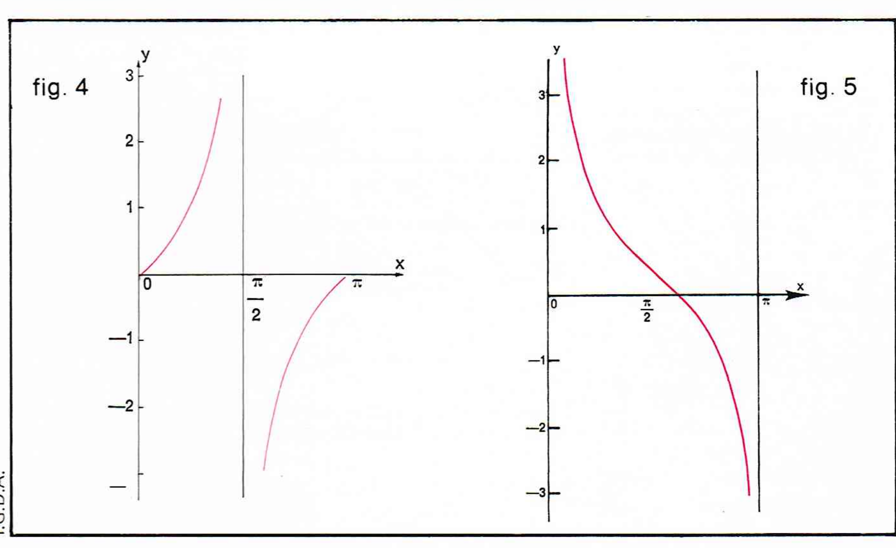
α	0	π/6	π/4	π/3	π/2
sin α	0	$\frac{1}{2}$	$\frac{\sqrt{2}}{2}$	$\frac{\sqrt{3}}{2}$	1
cos α	1	$\frac{\sqrt{3}}{2}$	$\frac{\sqrt{2}}{2}$	$\frac{1}{2}$	0
tg α	0	$\frac{\sqrt{3}}{3}$	1	$\sqrt{3}$	+ ∞
cotg α	+ ∞	$\sqrt{3}$	1	$\frac{\sqrt{3}}{3}$	0

▼ ***Figure* 3 :** *graphe des fonctions* sin x *et* cos x.
***Figure* 4 :** *graphe de la fonction* tg x.
***Figure* 5 :** *graphe de la fonction* cotg x.



x	0	π/2	π	3π/2	2π
cos x	1 →	0 →	−1 →	0 →	1
sin x	0 →	1 →	0 →	−1 →	0

Richard Colin



I.G.D.A.

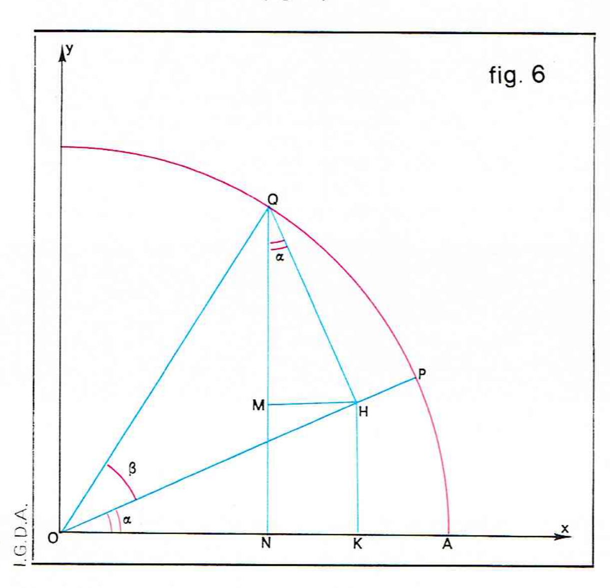
de ces fonctions, on peut toujours ajouter ou retrancher à leur argument un angle d'amplitude $2k\pi$ ($k \in \mathbb{N}$), et ainsi se ramener à un angle positif $< 2\pi$. On peut, d'autre part, toujours décomposer un angle quelconque β compris entre 0 et 2π en somme d'un angle α compris entre 0 et $\frac{\pi}{2}$ et d'un angle égal à $\frac{\pi}{2}$ $\left(\text{ou } \pi \text{ ou } \frac{3\pi}{2}\right)$, mais quand l'angle de départ est compris entre 0 et $-\frac{\pi}{2}$, on peut éviter de passer à un angle positif ; la réduction au premier quadrant s'effectue par l'intermédiaire du tableau suivant dont on vérifie facilement les résultats sur le cercle trigonométrique :

angle β (en radians)	sin β	cos β	tg β
π/2 − α	cos α	sin α	1/tg α
π/2 + α	cos α	− sin α	− 1/tg α
π − α	sin α	− cos α	− tg α
π + α	− sin α	− cos α	tg α
3π/2 − α	− cos α	− sin α	1/tg α
3π/2 + α	− cos α	sin α	− 1/tg α
− α	− sin α	cos α	− tg α

▶ *Tableau des principales* **fonctions trigonométriques.**

Formules d'addition et de soustraction

Calculons les fonctions circulaires des angles somme ou différence de deux angles α et β. Considérons le cercle trigonométrique et deux angles adjacents α et β d'origine OA ; soit $\overset{\frown}{AP}$ et $\overset{\frown}{PQ}$ les arcs correspondant à ces deux angles, N l'intersection de la perpendiculaire abaissée de Q sur OA, H l'intersection de la perpendiculaire abaissée de Q sur OP, M l'intersection avec QN de la parallèle à OA passant par H, et K l'intersection de la perpendiculaire abaissée de H sur OA (*fig.* 6).



▶ *Figure* **6** : *représentation graphique des* **formules d'addition.**

On a :

(1) $$\sin(\alpha+\beta) = \frac{NQ}{OA} = \frac{KH}{OA} + \frac{MQ}{OA} = \frac{OH}{OA}\sin\alpha + \frac{HQ}{OA}\cos\alpha = \sin\alpha\cos\beta + \cos\alpha\sin\beta\,;$$

(2) $$\cos(\alpha+\beta) = \frac{ON}{OA} = \frac{OK}{OA} - \frac{NK}{OA} = \frac{OH}{OA}\cos\alpha - \frac{HQ}{OA}\sin\alpha = \cos\alpha\cos\beta - \sin\alpha\sin\beta.$$

Ces deux formules constituent les **formules d'addition.** En remplaçant β par $-\beta$, on obtient les **formules de soustraction :**

(3) $\sin(\alpha-\beta) = \sin\alpha\cos\beta - \cos\alpha\sin\beta$

(4) $\cos(\alpha-\beta) = \cos\alpha\cos\beta + \sin\alpha\sin\beta$.

En posant $\beta = \alpha$, on obtient :

(5) $\sin 2\alpha = 2\sin\alpha\cos\alpha$;

(6) $\cos 2\alpha = \cos^2\alpha - \sin^2\alpha = 2\cos^2\alpha - 1 = 1 - 2\sin^2\alpha$.

En tenant compte de l'égalité $\text{tg}\,\alpha = \frac{\sin\alpha}{\cos\alpha}$ et des formules précédentes d'addition et de soustraction, on obtient :

(7) $$\text{tg}(\alpha+\beta) = \frac{\text{tg}\,\alpha + \text{tg}\,\beta}{1 - \text{tg}\,\alpha\,\text{tg}\,\beta}$$

(8) $$\text{tg}(\alpha-\beta) = \frac{\text{tg}\,\alpha - \text{tg}\,\beta}{1 + \text{tg}\,\alpha\,\text{tg}\,\beta}$$

(9) $$\text{tg}\,2\alpha = \frac{2\,\text{tg}\,\alpha}{1 - \text{tg}^2\alpha}$$

Dans de nombreux problèmes, on utilise les formules suivantes qui expriment $\sin\alpha$ et $\cos\alpha$ en fonction de la tangente de l'arc moitié : $\text{tg}\frac{\alpha}{2}$. On les obtient à partir des formules **(5)** et **(6)** en remplaçant α par $\frac{\alpha}{2}$, et enfin en divisant numérateur et dénominateur par $\cos^2\frac{\alpha}{2}$:

$$\sin\alpha = \frac{2\,\text{tg}\frac{\alpha}{2}}{1+\text{tg}^2\frac{\alpha}{2}} \qquad \cos\alpha = \frac{1-\text{tg}^2\frac{\alpha}{2}}{1+\text{tg}^2\frac{\alpha}{2}}$$

en divisant membre à membre, on obtient :

$$\text{tg}\,\alpha = \frac{2\,\text{tg}\frac{\alpha}{2}}{1-\text{tg}^2\frac{\alpha}{2}}$$

$\sin(\alpha+\beta) = \sin\alpha.\cos\beta + \cos\alpha.\sin\beta$

$\cos(\alpha+\beta) = \cos\alpha.\cos\beta - \sin\alpha.\sin\beta$

$\sin(\alpha-\beta) = \sin\alpha.\cos\beta - \cos\alpha.\sin\beta$

$\cos(\alpha-\beta) = \cos\alpha.\cos\beta + \sin\alpha.\sin\beta$

$$\text{tg}(\alpha+\beta) = \frac{\text{tg}\,\alpha + \text{tg}\,\beta}{1 - \text{tg}\,\alpha.\text{tg}\,\beta}$$

$$\text{tg}(\alpha-\beta) = \frac{\text{tg}\,\alpha - \text{tg}\,\beta}{1 + \text{tg}\,\alpha.\text{tg}\,\beta}$$

$\sin 2\alpha = 2\sin\alpha.\cos\alpha$

$\cos 2\alpha = \cos^2\alpha - \sin^2\alpha = 2\cos^2\alpha - 1 = 1 - 2\sin^2\alpha$

$$\text{tg}\,2\alpha = \frac{2\,\text{tg}\,\alpha}{1-\text{tg}^2\alpha}$$

$$\sin\alpha = \frac{2\text{tg}\frac{\alpha}{2}}{1+\text{tg}^2\frac{\alpha}{2}} \qquad \cos\alpha = \frac{1-\text{tg}^2\frac{\alpha}{2}}{1+\text{tg}^2\frac{\alpha}{2}}$$

$$\text{tg}\,\alpha = \frac{2\,\text{tg}\frac{\alpha}{2}}{1-\text{tg}^2\frac{\alpha}{2}}$$

Formules de transformation

On appelle ainsi les quatre formules qui transforment une somme ou une différence de sinus ou cosinus en un produit ; on les obtient en additionnant et soustrayant entre elles les formules **(1)** et **(3)**, et ensuite les formules **(2)** et **(4)**. On obtient ainsi *les formules de Werner :*



formules de Werner
$\sin\alpha\cos\beta = \frac{1}{2}\left[\sin(\alpha+\beta)+\sin(\alpha-\beta)\right]$
$\cos\alpha\sin\beta = \frac{1}{2}\left[\sin(\alpha+\beta)-\sin(\alpha-\beta)\right]$
$\cos\alpha\cos\beta = \frac{1}{2}\left[\cos(\alpha+\beta)+\cos(\alpha-\beta)\right]$
$\sin\alpha\sin\beta = -\frac{1}{2}\left[\cos(\alpha+\beta)-\cos(\alpha-\beta)\right]$

Posons $\alpha+\beta = p$ et $\alpha-\beta = q$, c'est-à-dire $\alpha = \left(\frac{p+q}{2}\right)$ et $\beta = \left(\frac{p-q}{2}\right)$, on obtient alors *les formules de transformation :*

formules de transformation
$\sin p + \sin q = 2\sin\frac{p+q}{2}\cos\frac{p-q}{2}$
$\sin p - \sin q = 2\cos\frac{p+q}{2}\sin\frac{p-q}{2}$
$\cos p + \cos q = 2\cos\frac{p+q}{2}\cos\frac{p-q}{2}$
$\cos p - \cos q = -2\sin\frac{p+q}{2}\sin\frac{p-q}{2}$

Fonctions trigonométriques circulaires inverses (ou réciproques)

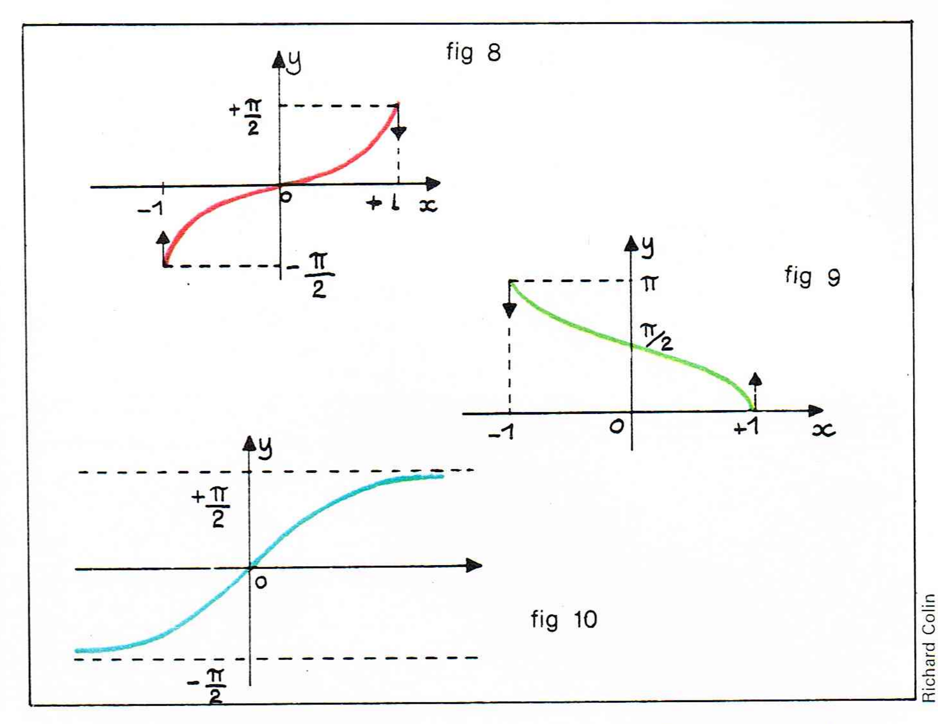
A toute valeur de la variable x, il correspond une valeur unique d'une fonction circulaire directe, mais la réciproque n'est pas vraie. En effet, considérons l'équation $\sin x = a$ où a est un nombre réel donné ; il y a une infinité de valeurs de x qui satisfont à cette équation si a est un nombre compris entre -1 et $+1$; sauf si $a = \pm 1$, il y a deux angles compris entre 0° et 360° qui satisfont l'équation. Considérons, par exemple, l'équation

$$\sin x = \frac{\sqrt{3}}{2},$$

prenons le point $\left(0, \frac{\sqrt{3}}{2}\right)$ et menons par ce point la parallèle à l'axe des x (*fig.* 7), on a deux intersections avec le cercle trigonométrique pour lesquelles l'équation est satisfaite :

$$x = \widehat{AOP} = \frac{\pi}{3}$$

$$x = \widehat{AOP'} = \frac{2\pi}{3}$$



Richard Colin

▲ ***Figure* 8 : *fonction* Arc sinus.**
***Figure* 9 : *fonction* Arc cosinus.**
***Figure* 10 : *fonction* Arc tangente.**

Il est pratique d'exprimer toutes les solutions d'une équation de ce type par une formule unique :

$$\sin x = a \Rightarrow x = (-1)^k\alpha + k\pi$$
$$\cos x = a \Rightarrow x = \pm\alpha + 2k\pi$$
$$\operatorname{tg} x = a \Rightarrow x = \alpha + k\pi$$

où α est le plus petit angle positif satisfaisant l'équation proposée, et k un entier relatif.

Lorsque les fonctions circulaires directes ont été introduites en analyse mathématique, il a été nécessaire de définir aussi leurs **fonctions réciproques.** Or, on a vu que ces fonctions ne sont pas monotones ; on ne pourra donc définir des fonctions réciproques que si on se restreint à des intervalles sur lesquels ces fonctions sont strictement monotones (voir *Analyse*).

On définit ainsi les fonctions **Arc sinus, Arc cosinus** et **Arc tangente** comme fonctions réciproques de la restriction du sinus à $\left[-\frac{\pi}{2}, +\frac{\pi}{2}\right]$, de la restriction du cosinus à $[0, \pi]$, et enfin de la restriction de la tangente à $\left[-\frac{\pi}{2}, +\frac{\pi}{2}\right]$ respectivement.

Ainsi, Arc sin est une bijection strictement croissante de $[-1, +1]$ sur $\left[-\frac{\pi}{2}, +\frac{\pi}{2}\right]$, et par définition (*fig.* 8) :

$$y = \operatorname{Arc\,sin} x \Leftrightarrow x = \sin y$$
$$-1 \leqslant x \leqslant +1 \qquad -\frac{\pi}{2} \leqslant y \leqslant +\frac{\pi}{2}$$

Arc cos est une bijection strictement décroissante de $[-1, +1]$ sur $[0, \pi]$ (*fig.* 9) :

$$y = \operatorname{Arc\,cos} x \Leftrightarrow x = \cos y$$
$$-1 \leqslant x \leqslant +1 \qquad 0 \leqslant y \leqslant \pi$$

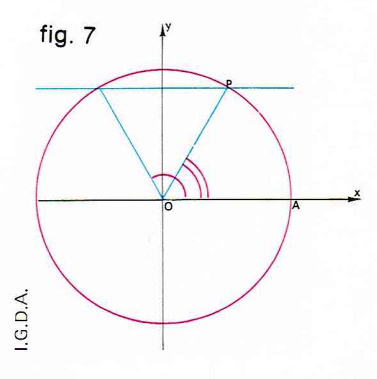
Arc tg est une bijection strictement croissante de $\mathbb{R}$ sur $\left]-\frac{\pi}{2}, +\frac{\pi}{2}\right[$ (*fig.* 10) :

$$y = \operatorname{Arc\,tg} x \Leftrightarrow x = \operatorname{tg} y$$
$$x \in \mathbb{R} \qquad -\frac{\pi}{2} \leqslant y \leqslant +\frac{\pi}{2}$$

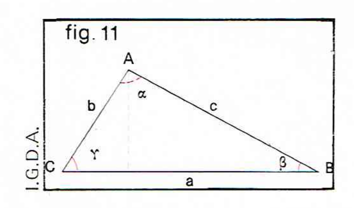
Le théorème des fonctions réciproques (voir *Analyse*) permet de calculer les dérivées de ces fonctions, on a :

$$(\operatorname{Arc\,sin} x)' = \frac{1}{\sqrt{1-x^2}} \text{ pour } x \in]-1, +1[$$

▼ ***Figure* 7 :** $\widehat{AOP} = \frac{\pi}{3}$ ***et*** $\widehat{AOP'} = \frac{2\pi}{3}$ ***sont les deux angles compris entre 0° et 360° qui satisfont l'équation*** $\sin x = \frac{\sqrt{3}}{2}$.



I.G.D.A.



▲ *Figure 11 : théorème des projections.*

$$(\text{Arc}\cos x)' = -\frac{1}{\sqrt{1-x^2}} \text{ pour } x \in]-1, +1[$$

$$(\text{Arc tg}\, x)' = \frac{1}{1+x^2}.$$

Ces fonctions sont très utilisées en calcul intégral et permettent d'écrire le nombre π comme intégrale définie ; par exemple : $\frac{\pi}{2} = \int_0^1 \frac{dx}{\sqrt{1-x^2}}$

Trigonométrie plane

La **trigonométrie plane** étudie les relations métriques entre les éléments de figures planes délimitées par des segments de droite. De telles figures étant toujours décomposables en triangles, la résolution des triangles est particulièrement importante. Dans ce paragraphe et dans le suivant, les angles seront mesurés en degrés.

Résoudre un triangle signifie : étant donné quelques éléments suffisant à le déterminer univoquement, calculer les autres. Un angle est généralement déterminé quand on connaît une de ses fonctions trigonométriques, car on sait qu'il est toujours < 180°, et dans le cas où on hésite entre un angle aigu et un angle obtus, on peut toujours recourir au calcul d'une seconde fonction du même angle.

Nous savons par la géométrie élémentaire qu'un triangle est déterminé quand trois de ses éléments, dont au moins un côté, sont donnés (un côté peut être remplacé par la surface).

Théorème des sinus

Quand trois éléments d'un triangle sont donnés, la solution s'obtient alors par la résolution d'un système de trois équations à trois inconnues. Comme première équation, on a celle de la géométrie élémentaire :

$$\alpha + \beta + \gamma = 180°$$

les deux autres sont données par le théorème des sinus qu'on admettra :

Dans un triangle quelconque, le rapport entre un côté et le sinus de l'angle opposé est constant et égal au diamètre du cercle circonscrit :

$$\frac{a}{\sin\alpha} = \frac{b}{\sin\beta} = \frac{c}{\sin\gamma}.$$

Théorème des projections

Tout côté d'un triangle est égal à la somme des produits de chacun des autres côtés par les cosinus des angles qu'ils forment avec le premier. Abaissons du sommet A d'un triangle la perpendiculaire sur le côté BC (*fig.* 11), on obtient deux triangles rectangles : si β et γ sont aigus, la base a est égale à la somme suivante :

$$a = b\cos\gamma + c\cos\beta$$

On peut vérifier que si γ (par exemple) est obtus, $\cos\gamma$ est négatif et la formule est toujours valable.

Théorème du cosinus ou de Carnot

Dans un triangle quelconque, le carré d'un des côtés est égal à la somme des carrés des autres diminuée du double produit de ces deux côtés par le cosinus de l'angle compris entre eux :

$$a^2 = b^2 + c^2 - 2\,bc\cos\alpha$$

Pour démontrer ce théorème, il suffit d'écrire les trois relations du théorème des projections relativement aux côtés a, b, c dans l'ordre, de multiplier respectivement par a, $-b$, $-c$, et puis d'additionner membre à membre.

Dans le cas particulier d'un triangle rectangle en A, l'expression se réduit au bien connu théorème de Pythagore.

On a encore en trigonométrie plane d'autres formules comme celles de **Neper** et de **Briggs** qui permettent d'exprimer une fonction trigonométrique d'un angle d'un triangle en fonction des autres angles et des côtés de ce triangle. Ces formules ont de multiples applications et permettent d'obtenir, en particulier, de nouvelles expressions de la surface d'un triangle comme la *formule d'Erone :*

$$S = \sqrt{p\,(p-a)\,(p-b)\,(p-c)},$$

où a, b, c sont les longueurs des trois côtés du triangle et $p = \frac{1}{2}(a+b+c)$.

Applications de la trigonométrie plane

Les fonctions trigonométriques et les formules de trigonométrie ont d'importantes et nombreuses applications. Les premières constituent un puissant moyen de calcul en mathématiques, en physique et dans toutes les sciences dont la théorie se base sur l'analyse mathématique, les secondes permettent de résoudre des problèmes pratiques de mesure, de distances et de calcul d'angles, en topographie, en géodésie, en astronomie.

Les fonctions trigonométriques sont évidemment très importantes pour la représentation des phénomènes périodiques. On peut écrire les développements de $\sin x$ et $\cos x$ en séries entières (voir *Analyse*) en exprimant x en radians :

$$\sin x = \sum_{n=0}^{+\infty} (-1)^{n+1} \cdot \frac{x^{2n+1}}{(2n+1)!}$$

$$\cos x = \sum_{n=0}^{+\infty} (-1)^n \frac{x^{2n}}{2n!}$$

Ces relations permettent de définir aussi les fonctions trigonométriques pour des valeurs complexes de x.

On peut aussi en déduire la *formule d'Euler :*

$$e^{ix} = \cos x + i\sin x.$$

Dans beaucoup d'ouvrages d'analyse, les fonctions trigonométriques sont introduites par la définition suivante : pour x réel, on appelle respectivement cosinus et sinus de x les parties réelles et imaginaire de e^{ix}, soit : $e^{ix} = \cos x + i\sin x$. Il en résulte que :

$$\cos t = \frac{e^{it} + e^{-it}}{2}$$

$$\sin t = \frac{e^{it} - e^{-it}}{2i}$$

et on déduit de ces formules et des propriétés de la fonction exponentielle les propriétés des fonctions trigonométriques.



▶ *Le mathématicien suisse* **Leonhard Euler** *(1707-1783).*

Trigonométrie hyperbolique

Donnons aussi les définitions des *fonctions hyperboliques* qui jouent pour la géométrie du *plan hyperbolique* le même rôle que les fonctions circulaires pour le *plan euclidien*. Pour tout x réel, on appelle *cosinus hyperbolique* de x, *sinus hyperbolique* de x, et *tangente hyperbolique* de x, respectivement les nombres :

$$\text{ch } x = \frac{1}{2}(e^x + e^{-x})$$

$$\text{sh } x = \frac{1}{2}(e^x - e^{-x})$$

$$\text{th } x = \frac{\text{sh } x}{\text{ch } x} = \frac{e^{2x} - 1}{e^{2x} + 1}.$$

Remarquons que le cosinus hyperbolique est une fonction paire, tandis que les deux autres sont des fonctions impaires. Un calcul simple montre que l'on a :

$$\text{ch}^2 x - \text{sh}^2 x = 1 \quad \text{pour tout } x.$$

Par dérivation, on obtient facilement :

$$(\text{ch } x)' = \text{sh } x,$$

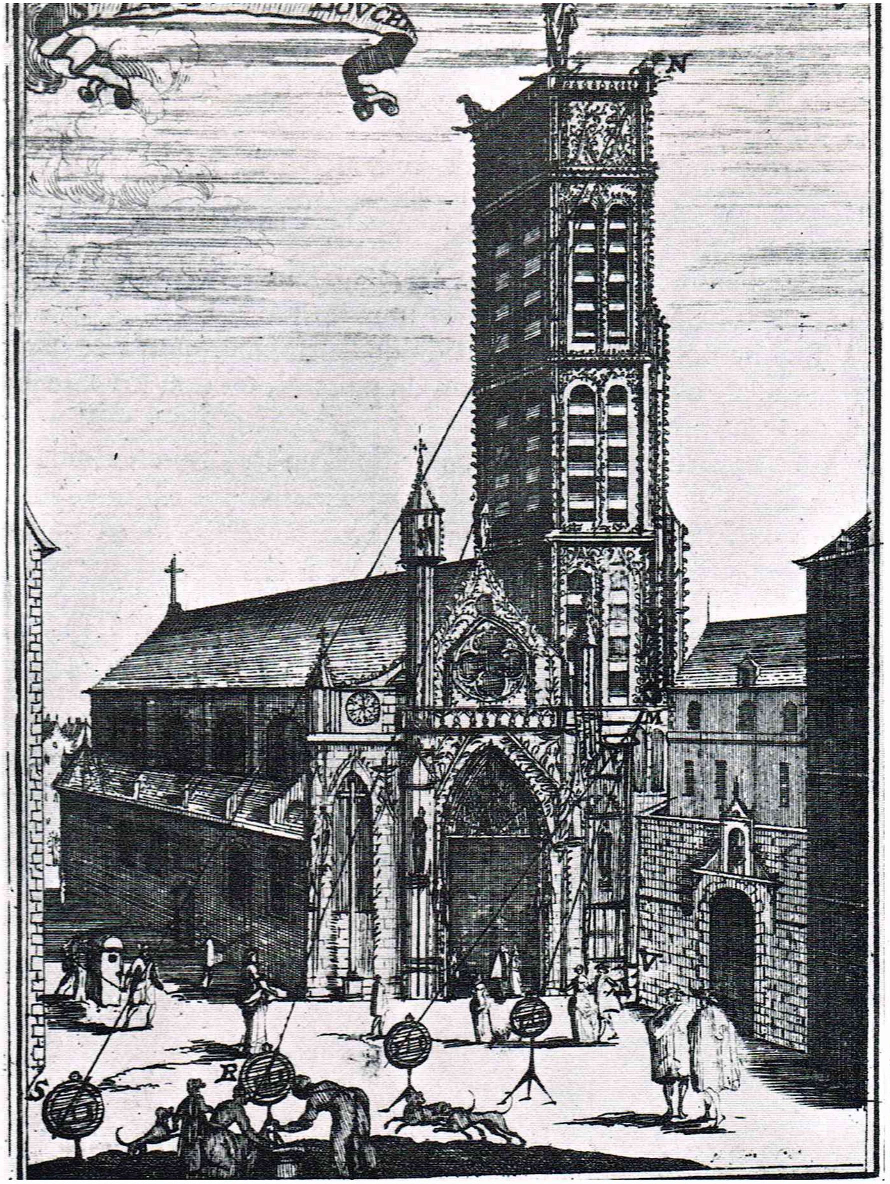
$$(\text{sh } x)' = \text{ch } x,$$

$$(\text{th } x)' = 1 - \text{th}^2 x = \frac{1}{\text{ch}^2 x}.$$

En considérant la définition de ch x, on s'aperçoit que ch x est toujours $\geqslant 1$, et il en résulte que sh x est une application strictement croissante de $\mathbb{R}$ dans $\mathbb{R}$. On voit que sh x tend vers $-\infty$ et $+\infty$ pour x tendant respectivement vers $-\infty$ et $+\infty$: cette fonction est donc bijective de $\mathbb{R}$ sur $\mathbb{R}$. Puisque sh $0 = 0$, le nombre sh x est du signe de x ; par suite, la fonction paire ch x est strictement croissante pour $x \geqslant 0$; d'autre part, ch x tend vers $+\infty$ quand x tend vers $+\infty$ ou vers $-\infty$. Enfin la fonction th x dont la dérivée est toujours strictement positive, est strictement croissante, et pour x tendant vers l'infini, on a :

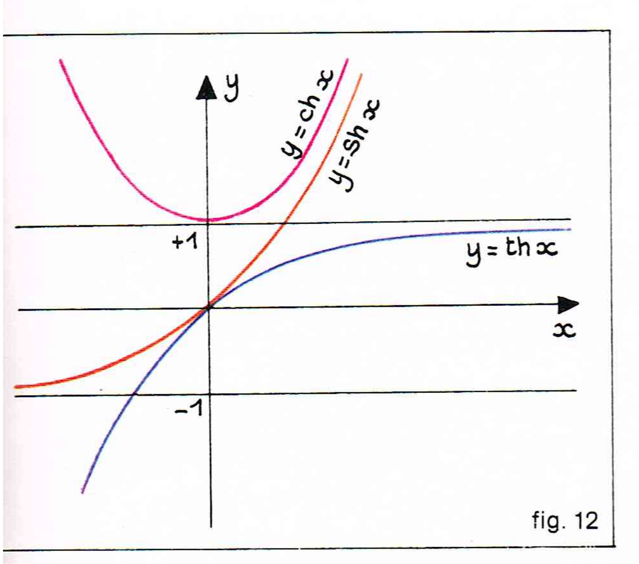
$$\lim_{x \to -\infty} \text{th } x = -1, \qquad \lim_{x \to +\infty} \text{th } x = +1 ;$$

on en déduit que th x est bijective de $\mathbb{R}$ sur $]-1, +1[$.



Collection Viollet

▲ ***Planche illustrant un manuel de trigonométrie du XVIIe siècle et montrant des calculs trigonométriques effectués à partir de l'église Saint-Jacques-la-Boucherie à Paris.***



◀ ***Figure 12 : les fonctions hyperboliques.***

On désigne par Arg sh x la bijection réciproque de $\mathbb{R}$ sur $\mathbb{R}$ de sh x, par Arg ch x la bijection réciproque de $[1, +\infty[$ sur $\mathbb{R}^+$ de ch x, et enfin par Arg th x la bijection réciproque de $]-1, +1[$ sur $\mathbb{R}$:

$$y = \text{Arg sh } x, x \in \mathbb{R} \Leftrightarrow x = \text{sh } y, \quad y \in \mathbb{R}$$

$$y = \text{Arg ch } x, x \geqslant 1 \Leftrightarrow x = \text{ch } y, \quad y \geqslant 0$$

$$y = \text{Arg th } x, -1 < x < +1 \Leftrightarrow x = \text{th } y, \quad y \in \mathbb{R}$$

Trigonométrie complexe

Les fonctions trigonométriques circulaires et hyperboliques s'étendent au domaine complexe de manière naturelle, en utilisant les développements en série ou (ce qui revient au même, puisque e^z est défini comme somme d'une série) au moyen des formules :

$$\cos z = \frac{e^{iz} + e^{-iz}}{2} \qquad \sin z = \frac{e^{iz} - e^{-iz}}{2i}$$

$$\text{ch } z = \frac{e^z + e^{-z}}{2} \qquad \text{sh } z = \frac{e^z - e^{-z}}{2}$$

qui mettent en évidence, par passage au domaine complexe, les liens étroits entre la trigonométrie circulaire et la trigonométrie hyperbolique. Ces fonctions sont analytiques dans tout le plan, et on a des relations du type :

$$\text{ch} z = \cos iz \qquad \sin iz = i \text{sh } z$$

qui sont valables en particulier quand z est un nombre réel et qui permettent de *déduire la trigonométrie hyperbolique de la trigonométrie circulaire et réciproquement.*

Applications pratiques

Pour les problèmes qui ont trait à la résolution des triangles, on mesure les angles en degrés. On va exposer quelques problèmes typiques dans lesquels interviennent des formules de trigonométrie : il sera ici simplement question de problèmes de topographie et d'astronomie où n'existent matériellement que les sommets des triangles. La mesure des angles se fait au moyen d'un théodolite.

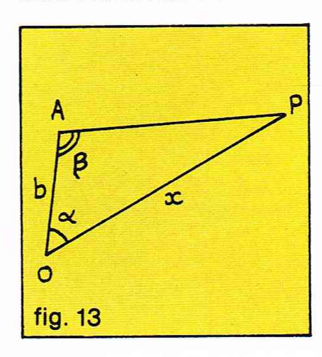
Dans la résolution des problèmes, on notera b la base qui est une longueur de référence dont la valeur est connue ; elle est choisie de telle sorte que le triangle dont elle fait partie soit celui dont on veut déterminer un élément.

Distance d'un point O à un point P visible, mais non accessible

On mesure la base AO $= b$ et les deux angles $\beta = \widehat{OAP}$ et $\alpha =$ AOP. Du théorème du sinus, on déduit (*fig.* 13) :

$$\text{OP} = \text{AO} \frac{\sin \beta}{\sin \widehat{\text{APO}}} = b \frac{\sin \beta}{\sin (\alpha + \beta)}$$

▼ ***Figure 13 : détermination de la distance d'un point O à un point P visible, mais non accessible.***

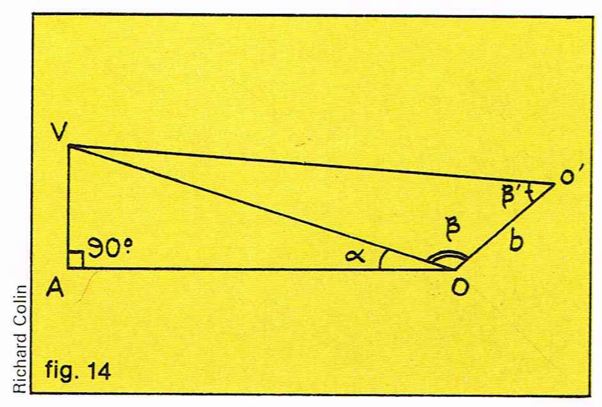


▶ *Page ci-contre, à gauche, figure* **18** *: théorème du cosinus; figure* **19** *: théorème du sinus.*

Il existe une méthode semblable pour mesurer la distance de deux points A et B visibles, mais non accessibles.

Hauteur d'une montagne relativement à un plan horizontal passant par un point O

Soit V le sommet de la montagne et A le pied de la perpendiculaire abaissée de V sur le plan horizontal passant par O (*fig.* 14).



▶ *Figure* **14** *: détermination de la hauteur d'une montagne relativement à un plan horizontal passant par un point* **O**.

On mesure la base OO′ = *b*, qui en général n'appartient pas à un plan horizontal, et les angles α, β, β′.

$$OV = b\,\frac{\sin\beta'}{\sin\widehat{OVO'}},\quad \widehat{OVO'} = 180° - (\beta + \beta')$$

$$AV = OV \sin\alpha$$

et en substituant, on obtient :

$$AV = b\,\frac{\sin\beta' \,\sin\alpha}{\sin(\beta+\beta')}$$

Distance de la Lune

On se trouve ici dans le cas du premier problème où P représente la Lune, et A et O deux points de la surface terrestre. On peut par exemple prendre comme base le rayon terrestre et lire la hauteur de la Lune en un point M′, celle-ci étant au zénith d'un point M″ situé sur le même méridien que M. Appliquons le théorème du sinus (*fig.* 15) :

$$\frac{R}{\sin\widehat{M'LO}} = \frac{OL}{\sin\widehat{LM'O}}$$

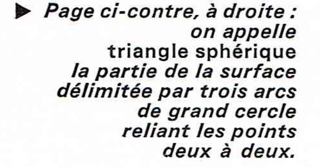
où R est le rayon terrestre; en tenant compte de

$$\widehat{OM'L} = 90° + \beta,$$

$$\widehat{M'LO} = 180° - (\alpha + \beta + 90°) = 90° - (\alpha + \beta),$$

on obtient :

$$OL = R\,\frac{\cos\beta}{\cos(\alpha+\beta)}.$$



▶ *Page ci-contre, à droite : on appelle* **triangle sphérique** *la partie de la surface délimitée par trois arcs de grand cercle reliant les points deux à deux.*

Distance du Soleil

Le problème ne peut pas se résoudre comme pour la Lune parce qu'une base quelconque de la Terre sera toujours trop petite par rapport à la distance à mesurer. Soit *l* la distance Terre-Lune. On mesure α quand la Lune est exactement à son premier quartier; la distance cherchée Terre-Soleil d_s est ainsi l'hypoténuse d'un triangle rectangle dont on connaît un côté et un angle (*fig.* 16) :

$$d_s = \frac{l}{\cos\alpha}$$

Il existe aussi une méthode pour calculer la distance des étoiles.

Trigonométrie sphérique

Considérons sur une surface sphérique de rayon R trois points n'appartenant pas au même grand cercle; on appelle **triangle sphérique** la partie de la surface délimitée par trois arcs de grand cercle (plus petits qu'un grand demi-cercle) reliant les points deux à deux (ceux-ci représentent sur la sphère leur distance minimale). La longueur des côtés du triangle sphérique est égale au produit du rayon de la sphère par l'angle au centre qui les sous-tend; l'**angle sphérique** est l'angle plan que font les deux tangentes aux deux côtés en leur point d'intersection.

Étant donné un côté d'un triangle sphérique, on peut, en le considérant comme un arc, parler d'une de ses fonctions trigonométriques. A tout angle d'un triangle sphérique, on peut faire correspondre un arc : on mène par le sommet de l'angle un diamètre de la sphère et le plan qui lui est perpendiculaire et qui passe par le centre de la sphère; ce dernier intersecte la sphère selon un grand cercle C qui, à son tour, rencontre les côtés de l'angle en deux points P et P′ (*fig.* 17). L'arc $\widehat{PP'}$ de C est l'arc cherché correspondant à l'angle α : puisque α′ = α, on a $\widehat{PP'} = \alpha \cdot R$. On n'en déduira pas que les éléments d'un triangle sphérique peuvent être considérés indifféremment arcs ou angles; dans ce paragraphe, on les mesurera en radians (en prenant le rayon comme unité de mesure, la mesure d'un côté donne aussi sa longueur).

La mesure de tout angle et de tout côté d'un triangle sphérique est comprise entre 0 *et* π.

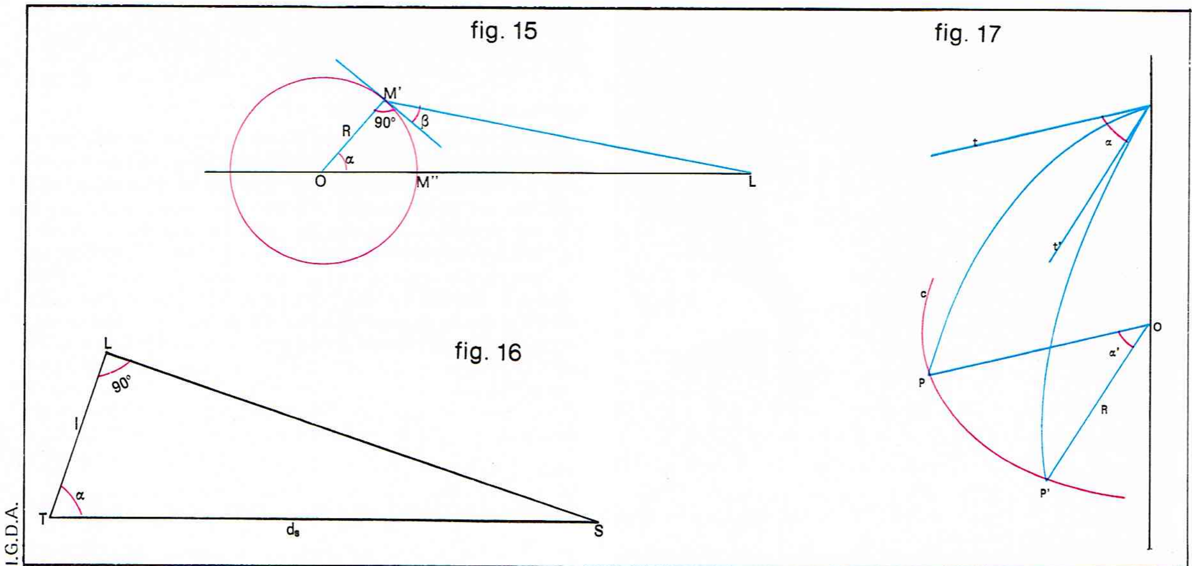
A tout triangle sphérique, on peut faire correspondre un *trièdre* de sommet O (centre de la sphère) et dont les arêtes passent par les sommets du triangle. Des propriétés du trièdre, on peut déduire celles du triangle sphérique :

— un côté est plus grand que la différence et plus petit que la somme des deux autres;

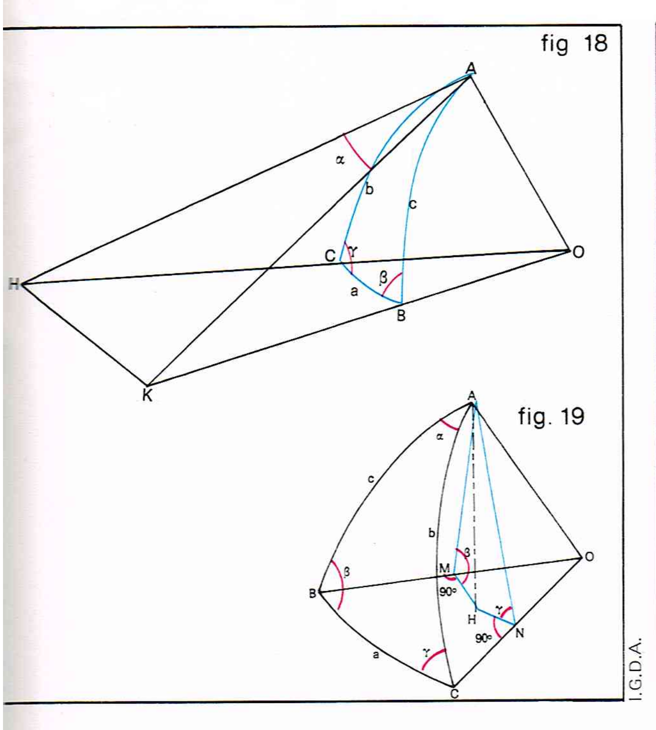
— la somme des côtés est comprise entre 0 et 2 π;

— la somme des angles est comprise entre π et 3 π;

— un angle externe est plus grand que la différence et plus petit que la somme des deux angles internes opposés :



▶ *Figure* **15** *: détermination de la distance de la Lune. Figure* **16** *: détermination de la distance du Soleil. Figure* **17** *: triangle sphérique.*



I.G.D.A.

— à des angles égaux, correspondent des côtés égaux ; au plus grand angle est opposé le plus grand côté.

On appelle **excès sphérique** d'un triangle la quantité dont la somme des angles du triangle sphérique dépasse celle des angles du triangle plan. En utilisant les notations introduites pour les triangles plans, on a :

$$e = \alpha + \beta + \gamma - \pi.$$

Pour la résolution des triangles sphériques, on peut établir des équations algébriques entre les fonctions circulaires des côtés et des angles.

Théorème du cosinus

On mène par un sommet, par exemple A, d'un triangle sphérique ABC, les tangentes aux deux côtés b et c (*fig.* 18) afin qu'elles rencontrent les deux autres arêtes du trièdre correspondant au triangle. En égalisant les deux expressions de HK obtenues avec le théorème de Carnot appliqué aux deux triangles plans HAK et HOK :

$$HK^2 = AH^2 + AK^2 - 2\,AH\,AK\cos\alpha$$

$$HK^2 = OH^2 + OK^2 - 2\,OH\,OK\cos a$$

$$AK = AO\,\mathrm{tg}\,c, \qquad AH = AO\,\mathrm{tg}\,b$$

$$OK = \frac{AO}{\cos c}, \qquad OH = \frac{AO}{\cos b}.$$

En égalisant et en substituant, on obtient :

$$\cos a = \cos b \cos c + \sin b \sin c \cos\alpha.$$

En écrivant les deux formules analogues, on a les systèmes suivants qui permettent de calculer trois éléments d'un triangle, connaissant les trois autres :

$$\begin{cases} \cos a = \cos b \cos c + \sin b \sin c \cos\alpha \\ \cos b = \cos c \cos a + \sin c \sin a \cos\beta \\ \cos c = \cos a \cos b + \sin a \sin b \cos\gamma \end{cases}$$

Théorème du sinus

En partant toujours du triangle sphérique ABC, on mène de A la perpendiculaire AH au plan BOC et de H la perpendiculaire aux rayons OB et OC (*fig.* 19). En tenant compte que par construction $\widehat{ANH} = \gamma$ et $\widehat{AMH} = \beta$ (AM et AN sont perpendiculaires respectivement à OB et OC), on a :

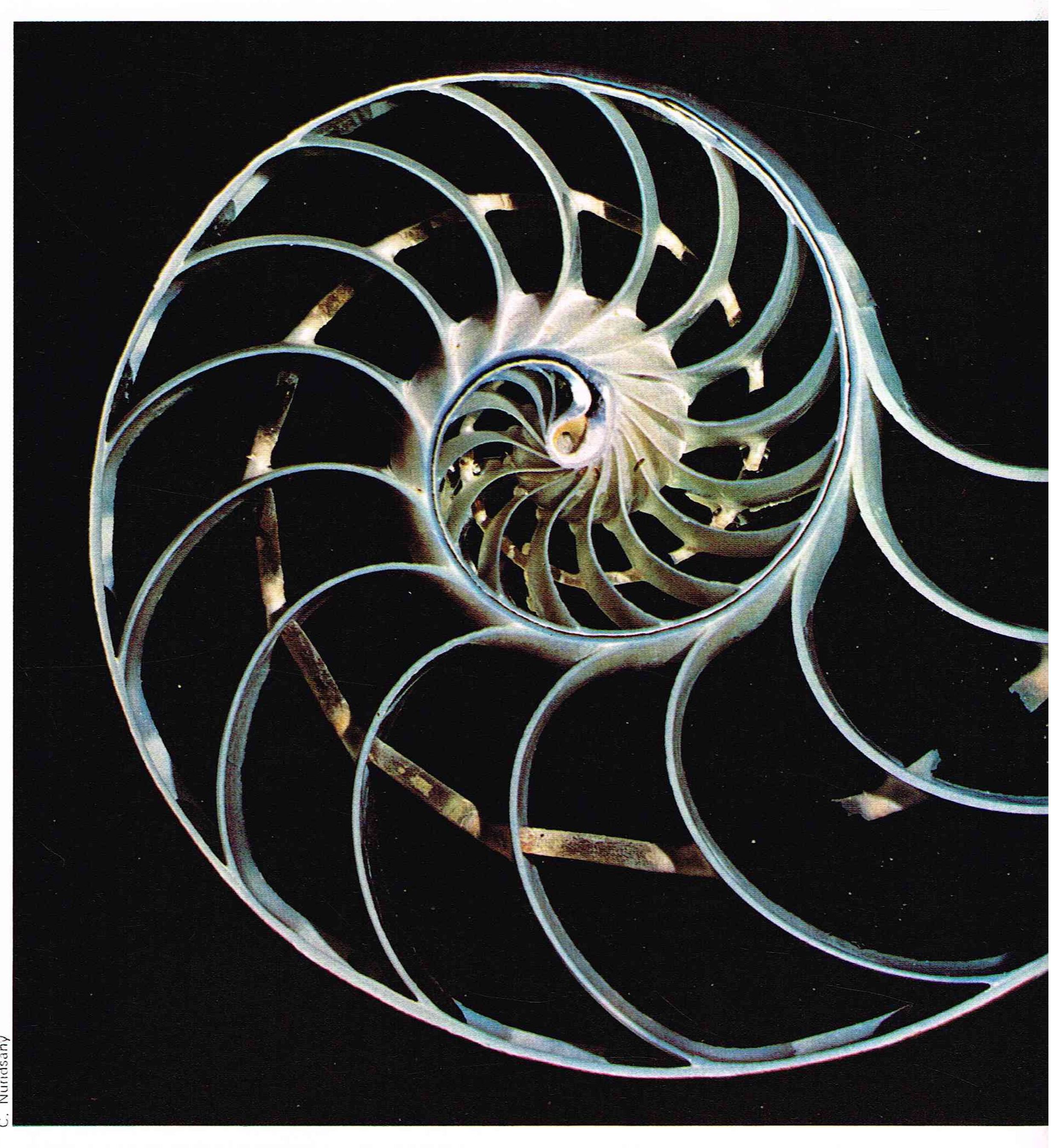
$$AH = AM \sin\beta = R \sin c \sin\beta$$

$$AH = AN \sin\gamma = R \sin b \sin\gamma.$$

En égalisant, on obtient :

$$\sin c \sin\beta = \sin b \sin\gamma$$

E. Rousseau - Snark International



C. Nuridsany

▶ *La coupe dans une coquille de nautile montre la succession des loges qui constituent une spirale à angle constant (spirale logarithmique).*

et l'énoncé du théorème est : *dans tout triangle sphérique, les sinus des côtés sont proportionnels aux sinus des angles opposés :*

$$\frac{\sin a}{\sin \alpha} = \frac{\sin b}{\sin \beta} = \frac{\sin c}{\sin \gamma}$$

A partir des formules des théorèmes du cosinus et du sinus, il est possible de résoudre un triangle sphérique ; comme il a déjà été vu avec les triangles plans, il est préférable d'utiliser aussi pour les triangles sphériques des formules qui permettent de déterminer immédiatement quelque élément inconnu. On tire de ces théorèmes l'expression des plus connues, comme celles de Delambre et Neper.

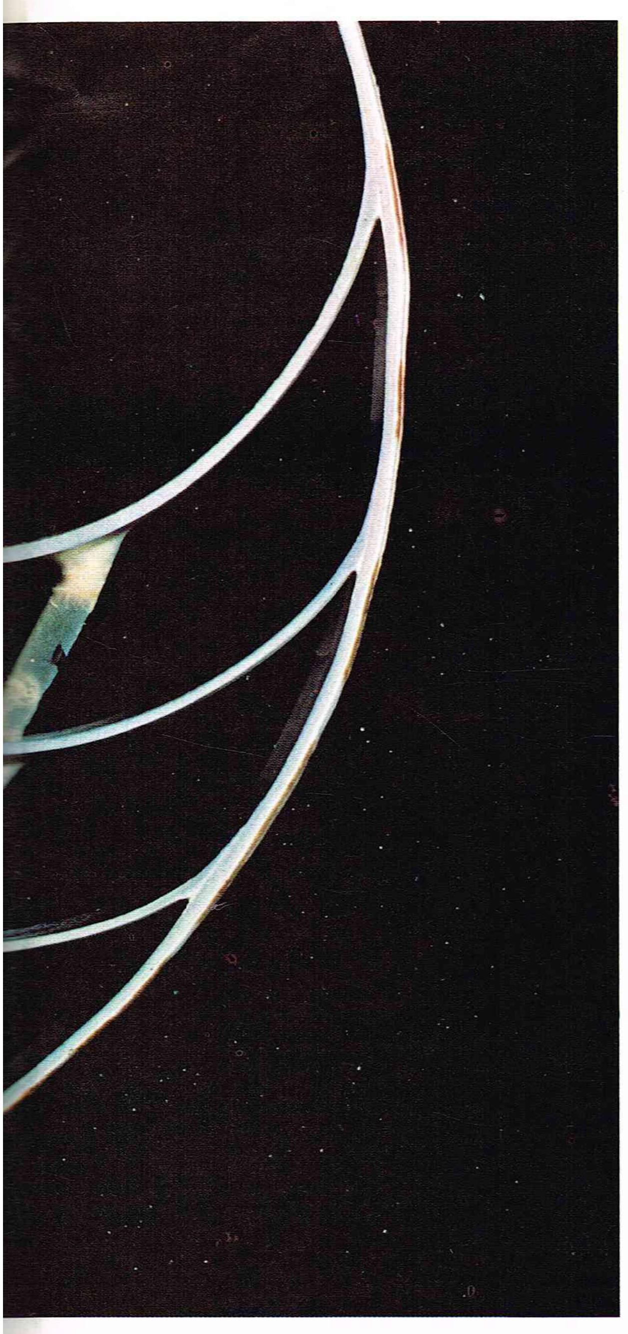
BIBLIOGRAPHIE

CAMPBELL R., *la Trigonométrie*, collection « Que sais-je ? », n° 692, Paris, 1956. - CHAMBADAL L. et OVAERT J.-L., *Cours de mathématiques, Notions fondamentales d'algèbre et d'analyse*, Gauthier-Villars, Paris, 1966. - DELACHET A., *les Logarithmes et leurs applications*, collection « Que sais-je ? », n° 850, Paris, 1970.

GÉOMÉTRIE ANALYTIQUE

La géométrie analytique permet d'étudier, à l'aide de l'analyse, les problèmes géométriques. C'est l'expression d'une réalité géométrique par une relation entre des variables grâce à l'usage d'un système de référence, ou système de coordonnées, et au principe de la représentation graphique.

Dans l'Antiquité, les coordonnées étaient surtout utilisées pour permettre des représentations astronomiques ou topographiques, mais déjà les Grecs, Archimède et surtout Apollonios, s'aperçurent de la possibilité de donner une forme graphique à des équations : Apollonios écrit explicitement les équations des coniques en coordonnées obliques. Plus tard, les développements de l'algèbre furent accompagnés de tentatives de résolution par l'intermédiaire d'interprétations géométriques, et la géométrie analytique, sous une forme encore très rudimentaire, fut utilisée avec un intérêt croissant par diverses branches mathématiques naissantes. Fermat et surtout Descartes, avec son œuvre *Géométrie* en 1637, lui fourniront un apport fondamental ; c'est au XVIII[e] siècle qu'elle prendra son essor et la forme qu'on lui connaît aujourd'hui. Elle s'étend alors à l'espace, et tous les grands mathématiciens de l'époque contribueront à son développement :



citons en particulier Euler, Lagrange et Monge. Outre les problèmes qu'elle permet de résoudre, la géométrie analytique a contribué à la naissance et au développement de nombreuses recherches, dans diverses branches des mathématiques, qui allaient se révéler extrêmement fécondes.

L'étude analytique d'un problème géométrique peut se séparer en trois phases :

— traduction des liens géométriques entre les éléments de la figure en expressions algébriques ;

— résolution purement analytique ;

— interprétation géométrique des résultats obtenus.

Nous allons voir ici les notions fondamentales qui permettent d'aborder une résolution en général et nous allons traiter, dans le plan et dans l'espace, les problèmes géométriques d'intérêt majeur. (Pour des développements ultérieurs, on pourra se reporter à la *Topologie*.)

Géométrie analytique plane

Coordonnées cartésiennes

Choisissons dans un plan deux droites orientées $x'Ox$ et $y'Oy$ (les *axes*) qui se coupent en O *(origine)*, et fixons sur chacune un vecteur unitaire, $\vec{i}$ et $\vec{j}$: on a construit un système de référence cartésien. Si P_1 est un point du plan, traçons par P_1 les parallèles aux axes ; chacune rencontre l'autre axe en un point H et K (*fig.* 1). Les coordonnées cartésiennes (x_1, y_1) de P_1 sont les mesures des segments OH, OK ; x_1 est l'*abscisse* de P_1, y_1 son *ordonnée*. Réciproquement, le couple (x_1, y_1) étant donné, l'intersection des droites parallèles à Oy et Ox menées par le point H d'abscisse x_1 sur Ox et le point K d'abscisse y_1 sur Oy est un point P_1 unique. On a donc une correspondance biunivoque entre l'ensemble des couples (x_1, y_1) éléments de $\mathbb{R}^2$ et le plan rapporté à un système d'axes de coordonnées. A partir des points O et P_1, on peut définir le vecteur OP_1, et on peut écrire :

$$\overrightarrow{OP_1} = x_1\vec{i} + y_1\vec{j}$$

où x_1, y_1 sont les composantes du vecteur $\overrightarrow{OP_1}$.

Si P_1 (x_1, y_1) et P_2 (x_2, y_2) sont deux points du plan, le vecteur d'origine P_1 et d'extrémité P_2 est le vecteur

$$\overrightarrow{P_1P_2} = (x_2 - x_1)\,\vec{i} + (y_2 - y_1)\,\vec{j} = \overrightarrow{OP_2} - \overrightarrow{OP_1}$$

Si les unités de mesure sont égales, nous pouvons utiliser la formule de Carnot (voir *Trigonométrie*) pour calculer la longueur du vecteur $\overrightarrow{P_1P_2}$, qui est aussi la distance des points P_1 et P_2 : si α est l'angle formé par les axes Ox et Oy, on aura :

$$|\overrightarrow{P_1P_2}| = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 - 2\,(x_2 - x_1)\,(y_2 - y_1)\cos\alpha}$$

Outre des unités de mesure égales, on prend habituellement $\alpha = \frac{\pi}{2}$ et on obtient ainsi un *système cartésien orthonormé* (c'est ce que nous utiliserons par la suite) ; dans ce cas, les composantes d'un vecteur coïncident avec ses projections orthogonales sur les axes, $\cos\alpha = 0$, et la formule de la distance de deux points se réduit à l'expression du théorème de Pythagore.

Changements de systèmes de coordonnées

Pour résoudre certains problèmes, il peut être utile, et même parfois indispensable, de passer à un deuxième système de référence obtenu d'abord par translation de l'origine et ensuite par une rotation des axes (c'est le cas le plus fréquent de passage entre deux systèmes cartésiens orthonormés). Comme le montre la *figure* 2, on trouve facilement les nouvelles coordonnées X et Y du point P (x, y) en fonction des anciennes coordonnées, de la rotation et de la translation (qui amènent les axes Ox et Oy à coïncider avec les axes OX et OY) ; on aura :

$$X = x\cos\theta - y\sin\theta + X_0$$
$$Y = y\cos\theta + x\sin\theta + Y_0$$

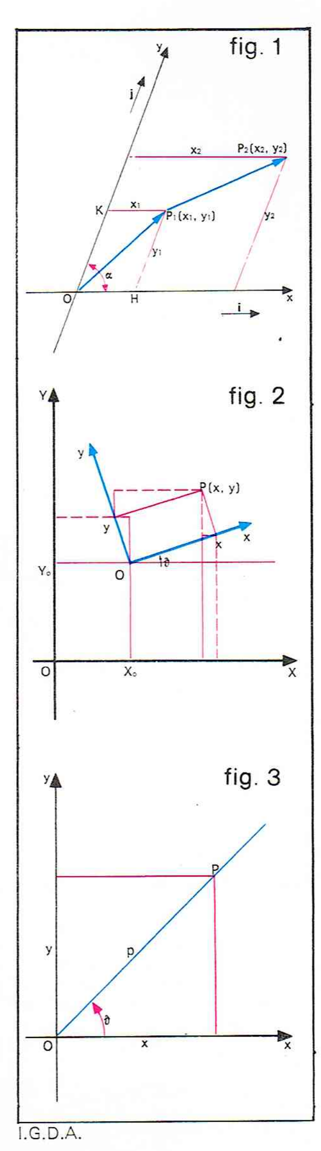
Autres types de coordonnées

Comme nous l'avons vu, un système de coordonnées dans un plan est une correspondance (biunivoque dans le cas cartésien) entre ce plan et $\mathbb{R}^2$; si l'on fixe l'une des coordonnées, on obtient une *courbe coordonnée :* dans le cas d'un repère cartésien, les courbes coordonnées sont des droites parallèles aux axes. Ce type de coordonnées est commode lorsqu'on a affaire à des figures polygonales (par exemple pour des calculs de longueurs ou d'aires). Mais dans d'autres cas, il sera de loin préférable de choisir un système dans lequel les courbes coordonnées sont du même type que la figure géométrique que nous voulons étudier. Nous allons voir ainsi trois autres types de coordonnées.

Coordonnées polaires

Fixons-nous une demi-droite orientée Ox de vecteur unitaire $\vec{i}$ (axe *polaire*) et d'origine le point O *(pôle)* ; les coordonnées polaires d'un point P sont : la distance $\rho = |\overrightarrow{OP}|$ *(rayon vecteur)* et θ l'angle $(\vec{i}, \overrightarrow{OP})$ *(angle polaire)* [*fig.* 3]. Pour passer des coordonnées polaires aux coordonnées cartésiennes d'un système orthonormé tel que l'axe polaire coïncide avec l'axe Ox et le pôle avec l'origine, on a immédiatement :

$$x = \rho\cos\theta \qquad \rho = \sqrt{x^2 + y^2}$$
$$y = \rho\sin\theta \qquad \theta = \text{Arc tg}\,\frac{y}{x}$$



I.G.D.A.

▲ *Figure* 1 : ***les coordonnées cartésiennes dans le plan.***
Figure 2 : ***changement du système de référence obtenu par translation de l'origine puis par rotation des axes.***
Figure 3 : ***les coordonnées polaires :*** ρ, ***rayon vecteur ;*** θ, ***l'angle polaire.***

▶ *Un autre type de construction complexe engendrée par la nature : une graine isolée d'Erodium ou « bec-de-héron » (Géraniacées) constitue une torsade spiralée.*

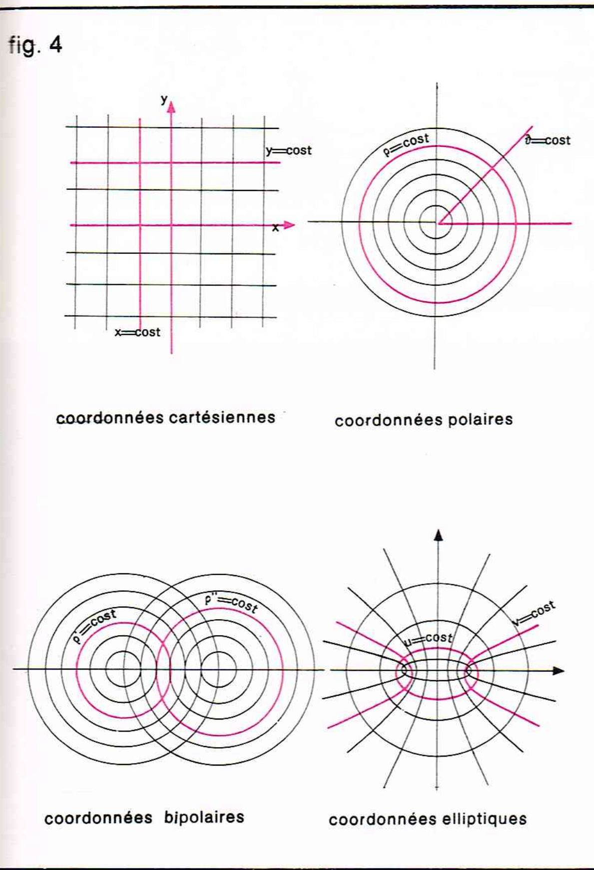


C. Nuridsany

Si on veut avoir une correspondance biunivoque entre les coordonnées polaires des points du plan et les points de ce plan (à l'exclusion du pôle), on doit ajouter les deux conditions : $0 < \rho < \infty$ et $0 \leqslant \theta < 2\pi$; le pôle est défini par la seule condition $\rho = 0$, θ étant arbitraire ; on l'appelle *point singulier* du système de coordonnées. Si, au contraire, on veut éviter la discontinuité qui résulte du pôle, on peut utiliser les coordonnées généralisées avec ρ et θ variant de $-\infty$ à $+\infty$ qui ne sont évidemment plus en correspondance biunivoque avec les points du plan. Les courbes coordonnées sont des cercles de centre O et des demi-droites d'origine O (*fig.* 4).

Coordonnées bipolaires

Soit F et F′ deux points fixes appelés *pôles* ou *foyers ;* les coordonnées bipolaires d'un point P sont les deux distances ρ et ρ' de P à F et F′. Il n'y a pas de correspondance biunivoque entre les coordonnées bipolaires et les points du plan puisque à chaque couple (ρ, ρ') (dont la somme est supérieure ou égale à la distance FF′) correspondent deux points qui sont symétriques par rapport à la droite passant par F et F′. Les courbes coordonnées sont des cercles ayant pour centre les deux foyers (*fig.* 4).



I.G.D.A.

◀ ***Figure 4 : coordonnées dans le plan et courbes coordonnées.***

Coordonnées elliptiques

Fixons-nous deux foyers F et F′ comme pour les coordonnées bipolaires ; les deux coordonnées elliptiques d'un point P sont respectivement la somme et la différence des deux coordonnées bipolaires :

$$u = \rho + \rho' \qquad v = \rho - \rho'$$

Les courbes coordonnées sont des ellipses et des hyperboles de foyers F et F′ (*fig.* 4).

La droite

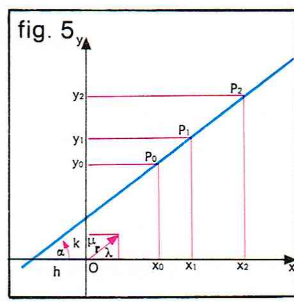
Fixons-nous un repère orthonormé dans le plan. La *droite* est le lieu géométrique caractérisé par la propriété suivante : le rapport des composantes de n'importe quel vecteur dont l'origine et l'extrémité sont sur la droite est constant. Considérant un vecteur ayant pour origine le point P_0 (x_0, y_0) sur une droite D et pour extrémité un point P (x, y) variable de D, on aura $\dfrac{(y - y_0)}{(x - x_0)} =$ constante. Notons que l'expression analytique d'une droite ou d'une courbe en général est constituée d'une (ou plusieurs) équation qui lie les coordonnées des points de la courbe : un point appartient à la courbe si et seulement si ses coordonnées vérifient l'équation (ou les équations simultanément). Si on choisit une valeur pour une des deux variables, l'équation permet en principe de trouver la ou les valeurs de l'autre.

Les composantes d'un vecteur libre unitaire parallèle à une droite,

$$\vec{r} = \lambda\vec{i} + \mu\vec{j}$$

sont appelées *cosinus directeurs* de la droite. Un vecteur libre parallèle à la droite sera un *vecteur directeur* de la droite. Les projections de $\vec{r}$, vecteur directeur unitaire, ne sont autres que les cosinus des deux angles que forme $\vec{r}$ avec $\vec{i}$ et $\vec{j}$, vecteurs unitaires des axes. Si on prend un vecteur $\overrightarrow{P_0P_1}$ constant et un vecteur $\overrightarrow{P_0P}$ variable tous les deux sur une droite donnée (*fig.* 5), on peut écrire :

$$\textbf{(1)} \qquad \frac{y - y_0}{x - x_0} = \frac{y_1 - y_0}{x_1 - x_0} = \frac{\mu}{\lambda} \qquad (\text{si } \lambda \neq 0)$$



I.G.D.A.

▲ ***Figure 5 : la droite dans le plan : le vecteur libre unitaire parallèle à la droite (r) et les cosinus directeurs (μ et λ).***

La première égalité fournit l'équation de la droite passant par deux points P_0 (x_0, y_0) et P_1 (x_1, y_1). Le rapport :

$$\frac{\mu}{\lambda} = m$$

est la tangente de l'angle formé par la droite avec l'axe Ox et appelé *coefficient angulaire* ou *pente* de la droite. En introduisant les symboles a, b, c, p, fonctions des constantes x_0, y_0, λ et μ, on obtient la forme :

$$ax + by + c = 0 \quad \text{ou} \quad y = mx + p$$

Si on veut mettre en évidence les segments h et k déterminés par l'intersection de la droite avec les axes de coordonnées, on peut écrire l'équation sous la forme :

$$\frac{x}{h} + \frac{y}{k} = 1$$

qui s'obtient facilement à partir de l'équation **(1)** en prenant comme points P_0 et P_1 les deux points d'intersection de la droite avec les axes.

Nous allons mentionner enfin les *équations paramétriques* de la droite qui nous donnent immédiatement les coordonnées, en fonction d'un paramètre t, des points d'une droite passant par P_0 (x_0, y_0) et de vecteur directeur $\vec{V}$ de composantes λ et μ :

$$x = x_0 + \lambda t$$
$$y = y_0 + \mu t$$

En éliminant t, on retrouve une des formules précédentes. Partant de **(1)**, on obtient les équations paramétriques en posant :

$$t = \frac{y - y_0}{\mu} = \frac{x - x_0}{\lambda}.$$

L'équation $y = mx + p$, avec m constante et p un paramètre variable, représente un *faisceau impropre* de droites, c'est-à-dire toutes les droites admettant pour vecteur directeur un vecteur de composantes λ et μ telles que $\dfrac{\mu}{\lambda} = m$. Plus généralement, l'équation du premier degré, combinaison linéaire de deux droites quelconques

$$\alpha\,(a_1x + b_1y + c_1) + \beta\,(a_2x + b_2y + c_2) = 0$$

représente toutes les droites d'un *faisceau*. Il est *propre* si les deux droites passent par un point M (l'équation est alors la représentation de toutes les droites passant par M), *impropre* lorsque les deux droites sont parallèles (voir *Géométrie*).

Pour trouver la condition de perpendicularité de deux droites, on peut rappeler l'expression du *produit scalaire* de deux vecteurs. Si

$$\vec{u} = a_1\vec{i} + b_1\vec{j} \quad \text{et} \quad \vec{v} = a_2\vec{i} \text{ et } b_2\vec{j}$$

sont un vecteur directeur de chacune des droites respectivement, les droites forment un angle α tel que

$$\vec{u} \cdot \vec{v} = |\vec{u}| \cdot |\vec{v}| \cdot \cos\alpha = a_1a_2 + b_1b_2$$

et lorsque α est un angle droit, on doit avoir

$$\vec{u} \cdot \vec{v} = a_1 a_2 + b_1 b_2 = 0 \quad \text{ou} \quad \frac{b_1}{a_1} = -\frac{a_2}{b_2}, \quad m_1 = -\frac{1}{m_2}$$

Il en résulte la règle suivante : une condition nécessaire et suffisante pour que deux droites soient perpendiculaires est que leurs coefficients angulaires vérifient la relation

$$m_1 m_2 = -1.$$

Le problème géométrique de l'intersection de deux droites se ramène au problème purement algébrique de la résolution d'un système de deux équations du premier degré à deux inconnues x et y. De façon analogue, le cas plus général de l'intersection de lieux géométriques représentés par des équations revient à la résolution du système formé par ces équations.

Pour trouver l'aire S d'un triangle de sommets $P_1\ (x_1, y_2)$, $P_2\ (x_2, y_2)$, $P_3\ (x_3, y_3)$, on a la formule suivante :

$$S = \frac{1}{2} \cdot \left| (x_2 - x_1)\ (y_3 - y_1) - (y_2 - y_1)\ (x_3 - x_1) \right|.$$

(En coordonnées obliques, il faut multiplier cette expression par sin α.)

Le cercle

L'équation d'un cercle dans un système orthonormé se trouve exprimée analytiquement comme le lieu des points situés à une distance constante d'un point fixé. Si on appelle C (α, β) le centre et r le rayon, un point sera à une distance r de C si et seulement si :

$$(x - \alpha)^2 + (y - \beta)^2 = r^2$$

qui devient :

$$x^2 + y^2 - 2\alpha x - 2\beta y + \gamma = 0 \quad (\gamma = \alpha^2 + \beta^2 - r^2).$$

On obtient ainsi une équation du second degré à deux inconnues qui présente les deux caractéristiques suivantes : les termes en x^2 et y^2 ont le même coefficient, et il n'y a pas de terme en xy. Étant donné une équation possédant ces deux caractéristiques :

$$ax^2 + ay^2 + bx + cy + d = 0 \quad (a \neq 0)$$

on peut encore l'écrire :

$$\left(x + \frac{b}{2a}\right)^2 + \left(y + \frac{c}{2a}\right)^2 + \frac{d}{a} - \left(\frac{b}{2a}\right)^2 - \left(\frac{c}{2a}\right)^2 = 0$$

Pour que cette équation soit celle d'un cercle de centre $\left(\alpha = -\frac{b}{2a}, \beta = -\frac{c}{2a}\right)$, de rayon r, il faut et il suffit que la valeur de r ainsi trouvée soit réelle :

$$r^2 = \left(\frac{b}{2a}\right)^2 + \left(\frac{c}{2a}\right)^2 - \frac{d}{a} \geqslant 0.$$

L'équation de la tangente à un cercle en un point $P_0\ (x_0, y_0)$ se déduit de la propriété d'orthogonalité entre le rayon et la tangente ; les cosinus directeurs du rayon CP_0 sont proportionnels à $x_0 - \alpha$ et $y_0 - \beta$, et l'équation de la tangente est (*fig.* 6) :

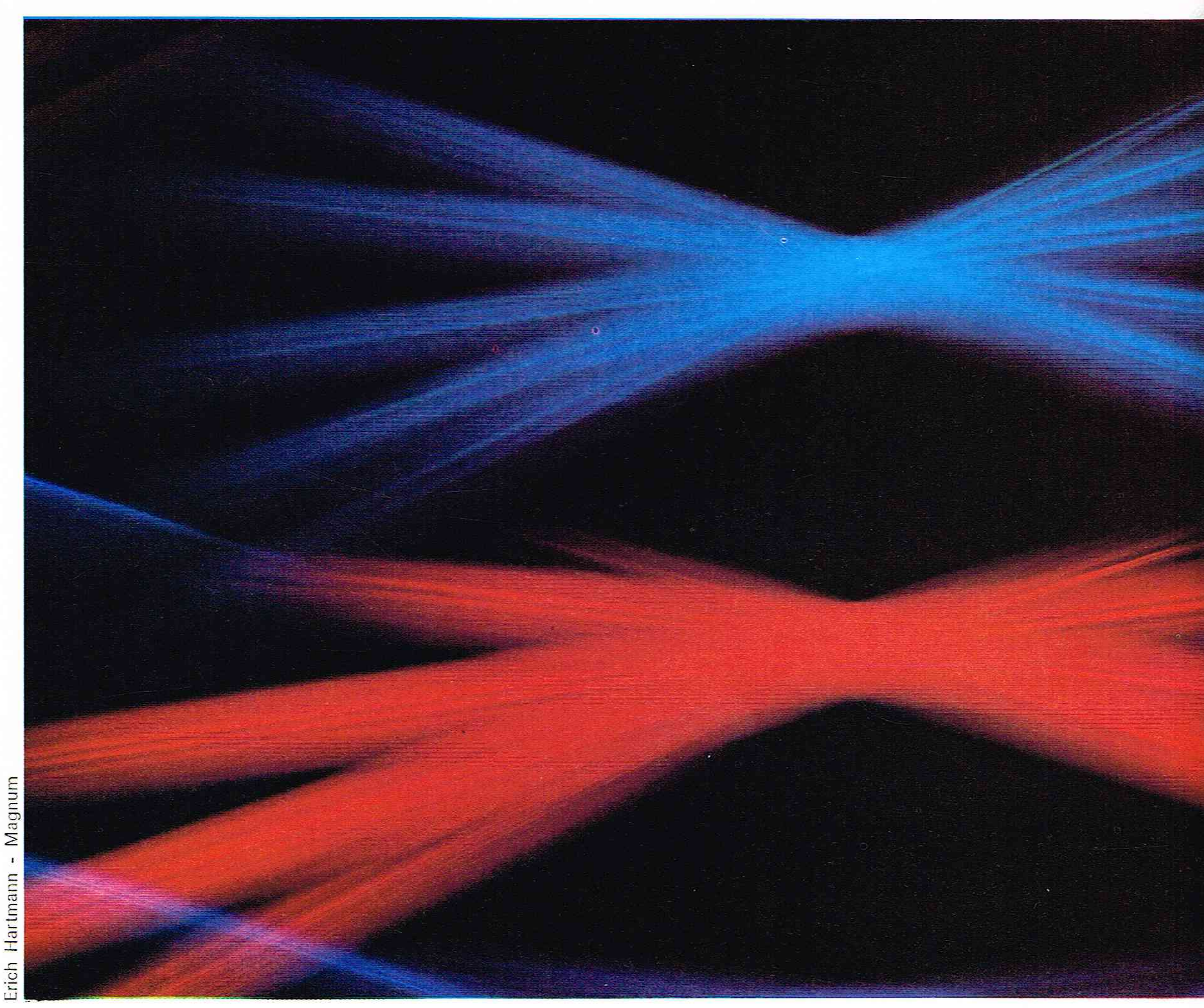
$$(x_0 - \alpha)\ (x - x_0) + (y_0 - \beta)\ (y - y_0) = 0$$

En coordonnées polaires, l'équation d'un cercle centré au pôle, de rayon r, se trouve considérablement simplifiée ; elle s'écrit en effet : $\rho = r$.

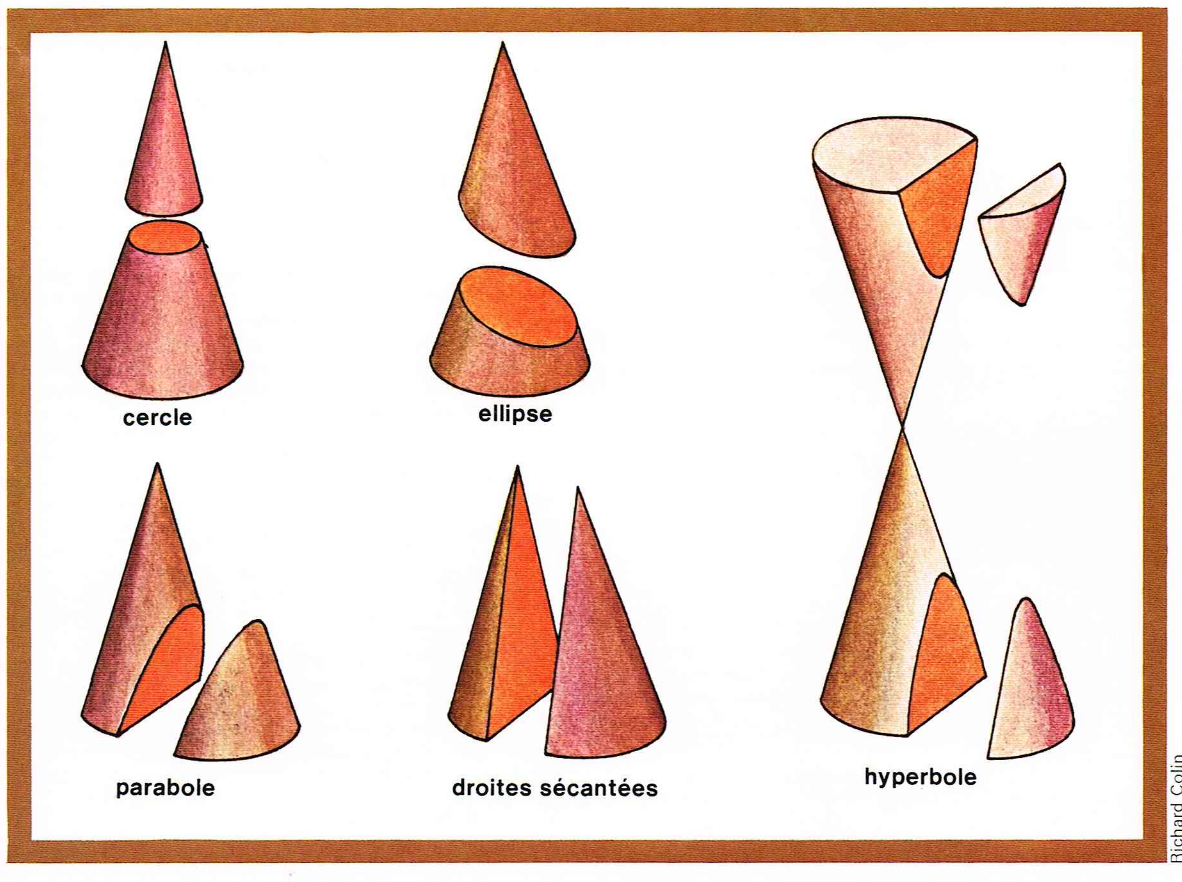
Les coniques

On désigne sous le nom de *coniques* les courbes obtenues par intersection d'un cône par un plan.

On se fixe : une droite f *(directrice)*, un point F *(foyer)* et un nombre réel positif e *(excentricité)* ; on appelle *conique propre* (ou non dégénérée) *réelle* le lieu des points P du plan qui satisfont la relation PF = r (constante) ou bien $\frac{PF}{PD} = e$ où D est la projection orthogonale

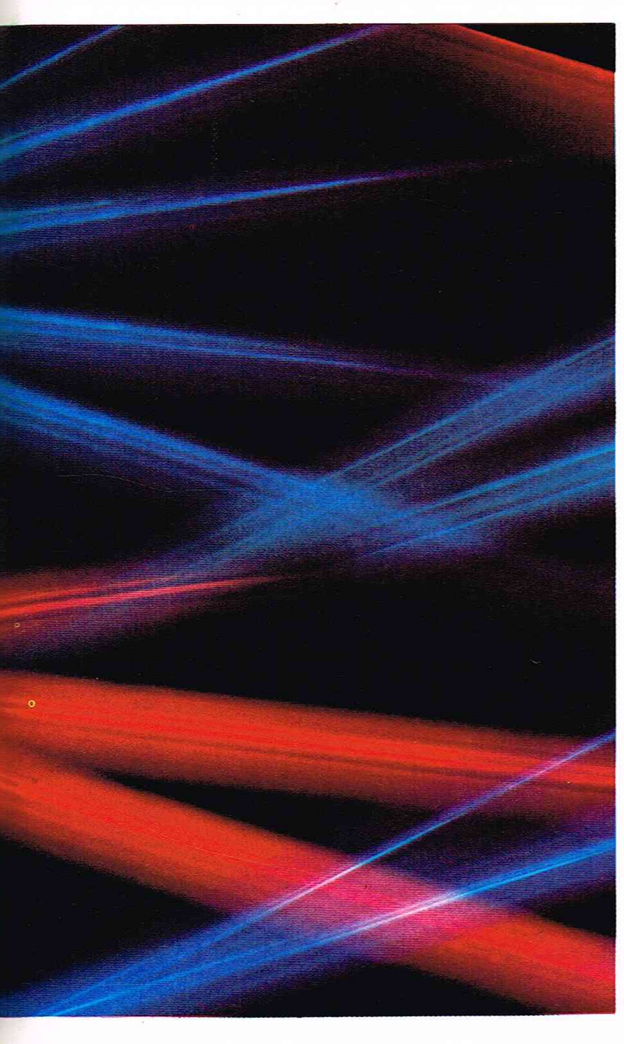


Erich Hartmann - Magnum



Richard Colin

◀ **Sections coniques d'Apollonios :** ***Apollonios découvrit que la section par un plan d'un cône à base circulaire engendre des courbes ou* coniques.**



de P sur *f*. Dans le premier cas, on a le cercle de centre F, de rayon *r*. Dans le second, on peut choisir un système de coordonnées (*fig.* 7) tel que l'axe O*x* coïncide avec la droite perpendiculaire à *f* passant par F. On appelle *c* et *h* respectivement les abscisses de F et D par rapport à une origine que nous fixerons tout à l'heure. La conique est le lieu des points dont les coordonnées sont les solutions de l'équation :

$$(1 - e^2)\, x^2 + y^2 - 2\,(c - e^2 h)\, x + c^2 - e^2 h^2 = 0.$$

Cette expression n'est pas la forme la plus générale de l'équation d'une conique à cause du choix particulier de l'axe O*x* : on démontrera cependant qu'une équation quelconque du second degré à deux inconnues se ramène à l'un des cas suivants :

a) il n'y a pas de solution réelle : dans ce cas, on parle de courbe *imaginaire ;*

b) le premier membre est le produit de deux polynômes linéaires, dans ce cas la conique est réduite à une ou deux droites : on parle de *conique dégénérée ;*

c) c'est l'équation d'une conique propre réelle.

Pour étudier les propriétés des coniques, nous allons choisir l'origine O du système de coordonnées de manière à rendre l'expression analytique de celles-ci la plus simple possible.

I) $e \neq 1$

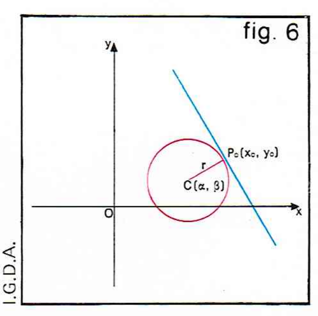
On choisira O tel qu'on ait l'égalité $\frac{c}{e^2} = h$ et on orientera l'axe O*x* de O vers F : l'équation peut alors s'écrire

$$\frac{x^2}{\frac{c^2}{e^2}} + \frac{y^2}{(1 - e^2)\,\frac{c^2}{e^2}} = 1$$

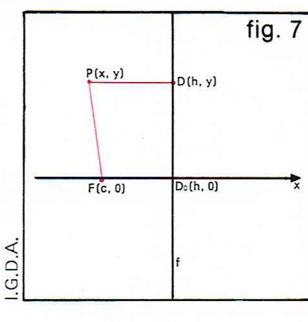
On peut poser $a = \frac{c}{e}$, $b^2 = \pm\,(1 - e^2)\,\frac{c^2}{e^2}$, et on devra considérer deux cas, selon que *e* est supérieur ou inférieur à 1 ; on obtient ainsi les deux types de coniques propres dont les équations peuvent s'écrire :

(1) $\qquad e < 1 \qquad \frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$

▼ ***Figure* 6 ; *le* cercle *de centre* C (α, β), *de rayon* r, *et sa tangente au point* P_0 (x_0, y_0).**



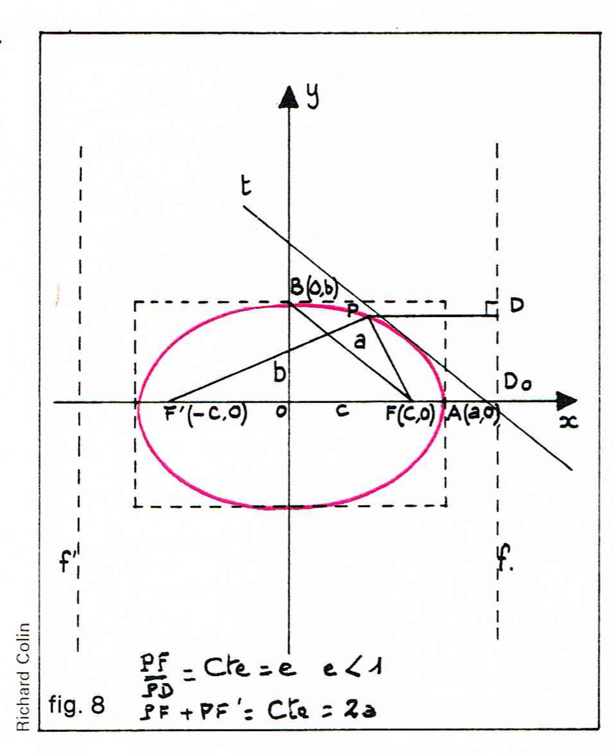
I.G.D.A.



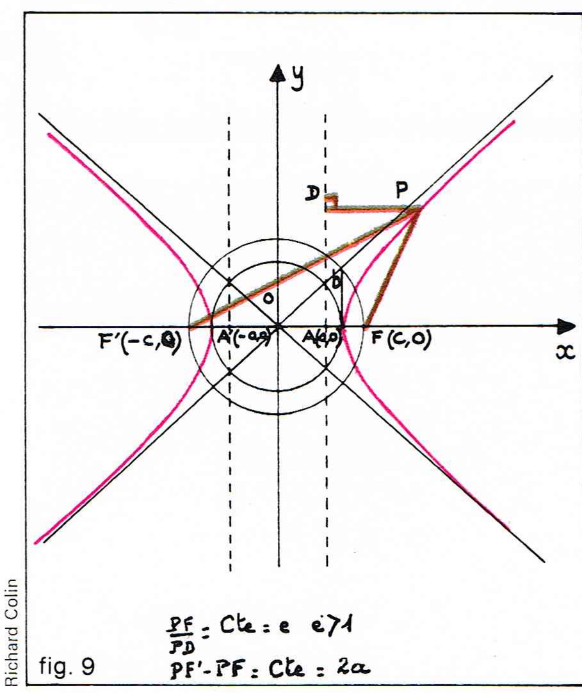
I.G.D.A.

▲ ***Figure* 7 ; *système de référence pour une* conique : *l'axe* x *coïncide avec la droite passant par le* foyer F *et perpendiculaire à la* directrice f.**

▶ ***Figure* 8 : *l'ellipse.***



▶ ***Figure* 9 : *l'hyperbole.***



(si $a = b$, on a un cercle)

(2) $\qquad e > 1 \qquad \dfrac{x^2}{a^2} - \dfrac{y^2}{b^2} = 1$

II) $e = 1$

Choisissons l'origine des axes au milieu de FD_0 ; on a ainsi :

(3) $\qquad y^2 - 4\,cx = 0$

Si l'équation d'une conique peut se mettre sous la forme **(1)**, on l'appelle une *ellipse*, sous la forme **(2)** une *hyperbole*, et sous la forme **(3)** une *parabole*.

Nous allons maintenant étudier séparément ces trois coniques.

L'ellipse

Considérons l'équation **(1)**. L'origine O, *centre* de l'ellipse, est centre de symétrie, les axes Ox et Oy (*axes* de l'ellipse) sont axes de symétrie. Si on explicite l'équation en x et en y, on trouve

$$y = \pm \frac{b}{a}\sqrt{a^2 - x^2} \qquad x = \pm \frac{a}{b}\sqrt{b^2 - y^2}$$

comme les expressions situées sous le radical doivent être positives, tous les points de la courbe sont situés à l'intérieur ou sur le rectangle centré à l'origine et de côtés 2 a et 2 b (*fig.* 8) ; les quatre points situés au milieu des côtés de ce rectangle sont les *sommets* de l'ellipse, a et b les *demi-axes*. En utilisant les deux égalités

$$a^2 = \frac{c^2}{e^2}, \qquad b^2 = (1 - e^2)\,\frac{c^2}{e^2},$$

on obtient l'expression de la distance focale en fonction des demi-axes : $c^2 = a^2 - b^2$; le foyer F est donc situé entre O et A. La directrice a pour équation : $x = \dfrac{c}{e^2} = \dfrac{a}{e}$ et le point D_0 est donc à l'extérieur de OA. Par symétrie de la courbe, on a un autre foyer F′ et une autre directrice f' ; l'axe Ox sur lequel sont situés les deux foyers est appelé *axe focal*.

L'équation de la tangente à l'ellipse en un point P_0 (x_0, y_0) est :

$$\frac{xx_0}{a^2} + \frac{yy_0}{b^2} = 1$$

Citons encore une propriété de l'ellipse : la somme des distances d'un point P de l'ellipse aux deux foyers est constante et égale à 2 a.

La représentation paramétrique de l'ellipse est :

$$x = a\cos\theta \qquad y = b\sin\theta$$

où θ est l'angle du vecteur $\overrightarrow{OP}$ avec l'axe Ox. En élevant au carré, en divisant la première égalité par a^2, la seconde par b^2 et en ajoutant, on retrouve bien l'équation **(1)**.

L'hyperbole

Revenons à l'équation **(2)**. De même que pour l'ellipse, l'origine O est centre de symétrie (*centre* de l'hyperbole), les axes de coordonnées sont des axes de symétrie. En explicitant l'équation, on obtient :

$$x = \pm \frac{a}{b}\sqrt{y^2 + b^2} \qquad y = \pm \frac{b}{a}\sqrt{x^2 - a^2}.$$

L'hyperbole passe donc par les deux points A $(a, 0)$ et A′ $(-a, 0)$ appelés *sommets* et se trouve à l'extérieur de la région du plan délimitée par les deux droites $x = -a$ et $x = a$ (elle ne rencontre donc pas l'axe Oy) ; elle admet comme *asymptotes* les deux droites $y = \pm \dfrac{b}{a}x$ dont elle se rapproche lorsque x tend vers l'infini, et est constituée de deux *branches*, comme on le voit sur la *figure* 9. Comme l'ellipse, elle a deux *foyers* F et F′ situés à l'extérieur du segment AA′ : des égalités

$$a^2 = \frac{c^2}{e^2}, \qquad b^2 = (e^2 - 1)\,\frac{c^2}{e^2}$$

on déduit leur distance à l'origine :

$$c = \sqrt{a^2 + b^2}.$$

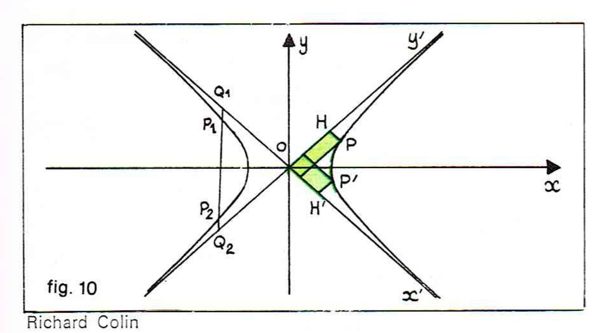
Les directrices ont pour équation $x = \pm \dfrac{a}{e}$, et comme $e > 1$, elles coupent le segment AA′ ; si l'on trace deux cercles centrés en O de rayons respectifs OA $= a$ et OF $= c$, le premier rencontre les asymptotes à leurs points d'intersection avec les directrices, le second les rencontre à leurs points d'intersection avec les droites $x = \pm a$. Il en résulte que b représente le segment de perpendiculaire à l'axe focal compris entre le sommet et une asymptote. La tangente à l'hyperbole en un point P_0 (x_0, y_0) a pour équation :

$$\frac{xx_0}{a^2} - \frac{yy_0}{b^2} = 1$$

Dans le cas particulier où $a = b$, les asymptotes sont perpendiculaires, et l'hyperbole est dite alors *équilatère*.

Si on prend comme nouveaux axes de coordonnées Ox' et Oy' les asymptotes (pas nécessairement perpendiculaires), l'équation de l'hyperbole prend une forme particulièrement simple :

$$x'y' = k^2 \left(k^2 = \frac{c^2}{4}\right).$$



Richard Colin

◀ *Figure 10 : deux propriétés de l'hyperbole : égalité des aires des parallélogrammes construits sur les asymptotes et ayant comme sommets opposés l'origine et un point* P *sur l'hyperbole ; égalité des segments* P_1Q_1 *et* P_2Q_2 *déterminés par l'intersection d'une corde avec l'hyperbole et les asymptotes.*

En utilisant cette équation de l'hyperbole, on en déduit que les parallélogrammes ayant deux côtés sur les asymptotes et comme sommets opposés l'origine O et un point P sur l'hyperbole ont une aire constante (*fig.* 10).

$A = x'y' \sin \alpha = k^2 \sin \alpha$ (α est l'angle formé par les axes de coordonnées).

Parmi les autres propriétés remarquables de l'hyperbole, citons les suivantes :

a) l'hyperbole est le lieu des points dont la différence des distances à deux points fixes (foyers) est constante ;

b) l'aire du triangle formé par une tangente à l'hyperbole et par les asymptotes est constante, quelle que soit la tangente ;

c) les deux segments P_1Q_1 et P_2Q_2 (*fig.* 10), déterminés par l'intersection d'une corde et des deux asymptotes, sont égaux ; cela rend très facile la construction de la courbe une fois connus les asymptotes et un point.

Les équations paramétriques de l'hyperbole s'expriment à l'aide des fonctions hyperboliques (voir *Trigonométrie*) :

$$x = a \operatorname{ch} t \qquad y = b \operatorname{sh} t \qquad \text{où } t \text{ est l'angle}$$

formé par $\overrightarrow{OP}$ avec l'axe Ox.

La parabole

Regardons maintenant l'équation **(3)** ; l'axe Ox (*axe de la parabole*) apparaît comme axe de symétrie. Elle admet un seul axe et un seul foyer F. Comme $e = 1$, on voit immédiatement que la parabole est le lieu des points équidistants d'une droite (la directrice) et d'un point (le foyer). La courbe passe par le milieu du segment FD_0, *sommet* de la parabole où elle est tangente à l'axe Oy (*fig.* 11). L'équation de la tangente en un point P_0 (x_0, y_0) est :

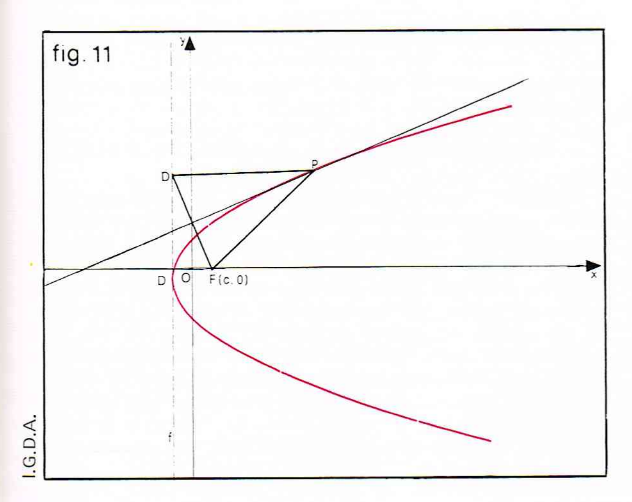
$$yy_0 - 2c\,(x + x_0) = 0$$

Parmi les propriétés remarquables de la parabole, retenons :

a) la directrice est le lieu des points symétriques du foyer par rapport aux tangentes, lorsqu'on fait varier celles-ci ;

b) le lieu des points qui sont la projection orthogonale du foyer sur la tangente, lorsque celle-ci varie, est la droite tangente à la parabole en son sommet (axe Oy) ;

c) la parabole est la limite d'une ellipse ou d'une hyperbole dont on a fait tendre un des foyers vers l'infini, l'autre restant fixe.



I.G.D.A.

◀ *Figure 11 : la* **parabole.**

Géométrie analytique dans l'espace

Coordonnées dans l'espace à trois dimensions

Dans l'espace, un système de coordonnées cartésiennes est formé de trois droites orientées (axes Ox, Oy, Oz), généralement orthogonales, passant par un point O (origine), munies de trois vecteurs unitaires $\vec{i}, \vec{j}, \vec{k}$, que nous supposerons de longueur égale. Les axes pris deux à deux déterminent les *plans de coordonnées ;* les *surfaces coordonnées* (par analogie avec les courbes coordonnées dans le plan) obtenues en fixant une coordonnée et en laissant varier les deux autres sont ici des plans perpendiculaires à l'axe dont la coordonnée est tenue constante. Par extension de ce qui se passe dans le plan, pour repérer le point P_1 de coordonnées x_1, y_1, z_1 (z_1 est la *cote*) et le vecteur $\overrightarrow{OP_1}$, on peut écrire :

$$P_1\,(x_1, y_1, z_1) \quad \overrightarrow{OP_1} = x_1\vec{i} + y_1\vec{j} + z_1\vec{k}$$

Si P_1 et P_2 sont deux points de l'espace, le vecteur d'origine P_1 et d'extrémité P_2 s'écrira :

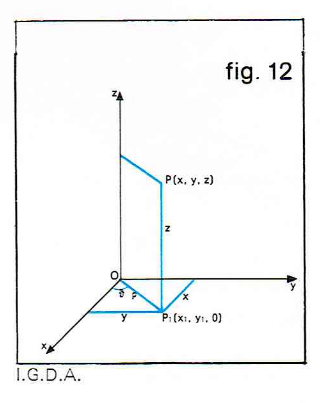
$$\overrightarrow{P_1P_2} = \overrightarrow{OP_2} - \overrightarrow{OP_1} = (x_2 - x_1)\,\vec{i} + (y_2 - y_1)\,\vec{j} + (z_2 - z_1)\,\vec{k}$$

et sa longueur, dans le cas d'un système orthonormé, sera :

$$\left|\overrightarrow{P_1P_2}\right| = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$$

En étendant à l'espace les coordonnées polaires par l'adjonction de l'axe normal Oz, on obtient les coordonnées *cylindriques* (ou *semi-polaires*) : les relations entre ρ, θ, x et y restent les mêmes, et la cote z est la même dans les deux systèmes (*fig.* 12). Dans ce cas, les surfaces coordonnées sont :

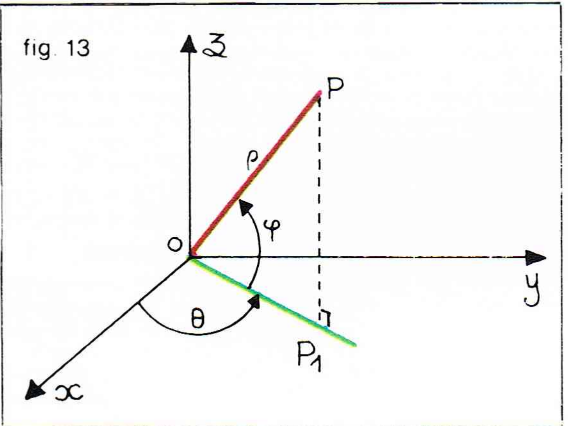
z = cste : plans perpendiculaires à l'axe Oz ;
θ = cste : demi-plans d'origine Oz et formant l'angle θ avec le plan xOz ;
ρ = cste : cylindre d'axe Oz et de rayon ρ.



I.G.D.A.

▼ **Figure 12 ; coordonnées cylindriques :** *ρ et θ sont les coordonnées polaires du plan et z coïncide avec la coordonnée cartésienne.*

Comme autre extension du système de coordonnées polaires, on a le système de coordonnées *sphériques :* les trois coordonnées sont alors (*fig.* 13) :



Richard Colin

◀ *Figure 13 :* **coordonnées sphériques.**

$\rho = \left|\overrightarrow{OP}\right|$ *rayon vecteur,* distance du point à l'origine ;

φ, *latitude*, est l'angle ($\overrightarrow{OP_1}$, $\overrightarrow{OP}$) où P_1 est la projection orthogonale de P sur le plan xOy ;

θ, *longitude*, est l'angle ($\overrightarrow{Ox}$, $\overrightarrow{OP_1}$).

Les relations entre les coordonnées sphériques et les coordonnées cartésiennes sont les suivantes :

$$x = \rho \cos \varphi \cos \theta ; \quad y = \rho \cos \varphi \sin \theta ; \quad z = \rho \sin \varphi ;$$

et les surfaces coordonnées :

ρ = cste : sphère de centre O de rayon ρ ;
φ = cste : demi-cône de révolution d'axe Oz ;
θ = cste : demi-plan d'origine l'axe Oz et formant l'angle φ avec le demi-axe positif Ox.

Choisissant un système dans lequel l'axe Ox coïncide avec l'axe terrestre et d'origine le centre de la Terre, en coupant les surfaces coordonnées par la surface de la Terre (supposée sphérique), on obtient les *parallèles* (φ = cste) et les *méridiens* (θ = cste).

▼ ***Ci-dessous, surfaces réglées. En bas, figure* 14 : cylindre *et* cône de révolution.**

La droite et le plan

L'espace étant rapporté à un système orthonormé, un plan peut être défini par un point et une direction perpendiculaire. La direction est donnée par un vecteur libre $\vec{u}$ de composantes *a*, *b*, *c*, et le point P_0 par ses coordonnées x_0, y_0, z_0. La condition que doivent vérifier tous les vecteurs du plan d'origine P_0 et d'extrémité un point quelconque du plan, P, est :

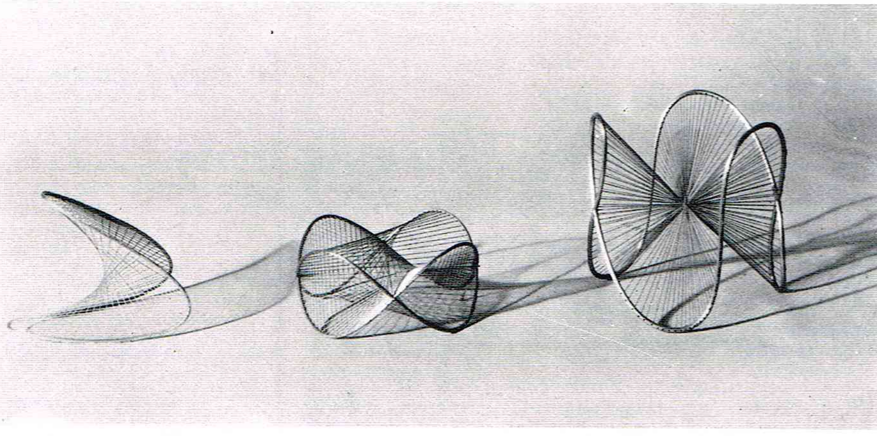
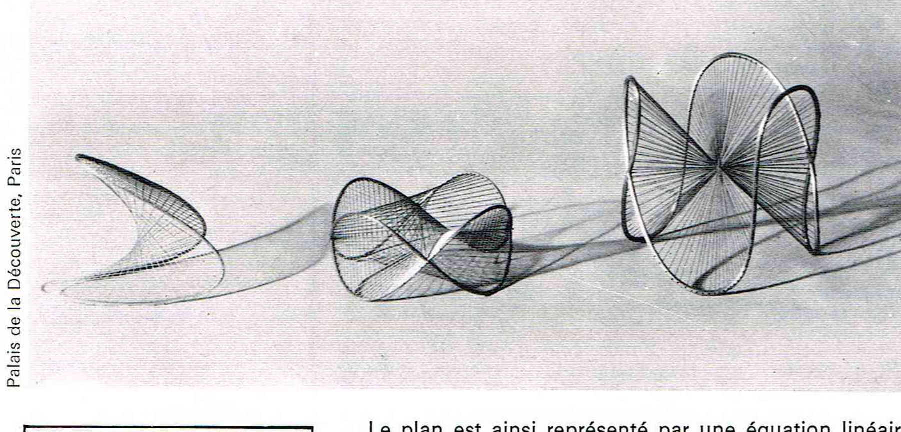
$$\overrightarrow{P_0P} \cdot \vec{u} = 0;$$

explicitant le produit scalaire, on obtient

$$a\,(x - x_0) + b\,(y - y_0) + c\,(z - z_0) = 0$$

et on trouve une équation de type

$$ax + by + cz + d = 0$$

Palais de la Découverte, Paris

fig. 14

I.G.D.A.

Le plan est ainsi représenté par une équation linéaire en *x*, *y* et *z*, et on peut voir que toute équation de ce type représente un plan perpendiculaire au vecteur de composantes *a*, *b*, *c* ; on peut remarquer que, si une variable ne figure pas dans l'équation d'un plan, celui-ci est parallèle à l'axe correspondant à la variable manquante. Si deux coefficients sont nuls, le plan est parallèle à l'un des trois plans de coordonnées.

Une droite peut être caractérisée comme l'intersection de deux plans, et on peut donc la représenter par le système formé des équations de deux plans qui la contiennent (*équations cartésiennes* de la droite). On peut aussi la définir (comme on l'a vu dans le plan) par un point P_0 (x_0, y_0, z_0) et un vecteur directeur $\vec{r}$ (*l*, *m*, *n*), et on obtient les « équations paramétriques » de la droite, où λ est un paramètre :

$$\begin{cases} x = x_0 + \lambda l \\ y = y_0 + \lambda m \\ z = z_0 + \lambda n \end{cases}$$

qui peuvent se mettre sous la forme (si les composantes de $\vec{r}$ sont non nulles)

$$\frac{(x - x_0)}{l} = \frac{(y - y_0)}{m} = \frac{(z - z_0)}{n} = \lambda$$

et on obtient alors la forme *normale* de l'équation d'une droite.

Si on se donne les équations de deux droites non parallèles

$$\frac{(x - x_1)}{l_1} = \frac{(y - y_1)}{m_1} = \frac{(z - z_1)}{n_1}$$

$$\frac{(x - x_2)}{l_2} = \frac{(y - y_2)}{m_2} = \frac{(z - z_2)}{n_2},$$

une condition nécessaire et suffisante pour qu'elles soient coplanaires est que leurs vecteurs directeurs $\vec{r_1}$ et $\vec{r_2}$ et le vecteur $\overrightarrow{P_1P_2}$ soient coplanaires [où P_1 (x_1, y_1, z_1) est sur la première, P_2 (x_2, y_2, z_2) sur la deuxième], c'est-à-dire :

$$\begin{vmatrix} x_2 - x_1 & l_1 & l_2 \\ y_2 - y_1 & m_1 & m_2 \\ z_2 - z_1 & n_1 & n_2 \end{vmatrix} = 0$$

Surfaces de révolution et surfaces réglées

On appelle *surface de révolution* la surface décrite par une courbe quelconque lorsqu'on la fait tourner autour d'une droite *(axe de révolution)* ; une surface est *réglée* si, en n'importe lequel de ses points, il passe au moins une droite tout entière contenue dans la surface.

La sphère

C'est le lieu des points situés à une distance *r* d'un point C (α, β, γ) qui est le *centre* de la sphère ; les coordonnées de ses points vérifient l'équation :

$$(x - \alpha)^2 + (y - \beta)^2 + (z - \gamma)^2 - r^2 = 0$$

L'intersection d'une sphère avec un plan est soit vide, soit un point (plan tangent), soit un cercle.

Le cône et le cylindre

Étant donné deux droites coplanaires, si l'on fait tourner l'une d'elles autour de l'autre (axe de révolution) on obtiendra une surface qui sera un *cône de révolution* si les droites se coupent en un point *(sommet)* ou un *cylindre de révolution* si les droites sont parallèles (*fig.* 14). Plus généralement, on appelle *cône* la surface décrite par toutes les droites *(génératrices)* passant par un point fixe et par un point d'une courbe $\mathcal{L}$ *(directrice)*, et *cylindre* celle décrite par toutes les droites parallèles passant par un point de $\mathcal{L}$; ce sont évidemment, de par leur construction, des surfaces réglées. L'équation d'un cylindre de révolution d'axe O*z* et de rayon *r* est :

$$x^2 + y^2 = r^2.$$

Celle d'un cône de révolution d'axe O*z*, de sommet l'origine, est :

$$x^2 + y^2 - m^2 z^2 = 0.$$

Les intersections avec des plans $z = h$ donnent des cercles (de rayon *r* pour le cylindre) centrés sur l'axe O*z*, alors que les intersections avec des plans contenant l'axe O*z* seront des couples de droites.

Les quadriques

On appelle *quadrique* toute surface dont l'équation en coordonnées cartésiennes est du second degré en *x*, *y* et *z*. C'est ce qui correspond, dans l'espace, aux coniques dans le plan. Les intersections des quadriques avec les plans de coordonnées et des plans parallèles à ceux-ci sont des coniques : une fois celles-ci connues, on connaît le type et l'allure de la surface. On peut considérer toutes les quadriques comme les surfaces décrites par des coniques situées dans des plans parallèles lorsque leurs sommets décrivent deux autres coniques. Nous avons les types suivants :

Ellipsoïde

C'est une quadrique qui peut se représenter par l'équation suivante :

$$\frac{x^2}{a^2} + \frac{y^2}{b_2} + \frac{z^2}{c^2} = 1$$

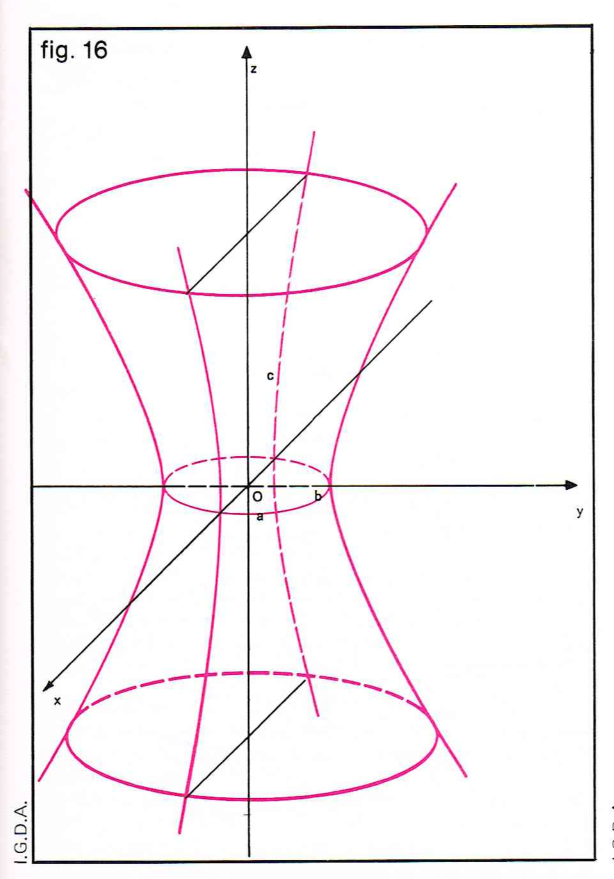
Les trois plans de coordonnées la coupent suivant les ellipses principales (*fig.* 15)

plan $z = 0 : \frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ (de demi-axes *a* et *b*)

plan $y = 0 : \frac{x^2}{a^2} + \frac{z^2}{c^2} = 1$ (de demi-axes *a* et *c*)

plan $x = 0 : \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1$ (de demi-axes *b* et *c*)

Les plans de coordonnées sont des plans de symétrie de l'ellipse, les axes de coordonnées sont des axes de symétrie. Tous les points de l'ellipsoïde sont à l'intérieur du parallélépipède de côtés 2 *a*, 2 *b* et 2 *c*, et toutes les sections pour des plans sont des ellipses. Si deux demi-axes sont égaux (par exemple, $a = b$), on a un ellipsoïde de révolution (ici autour de l'axe O*z*) ; dans ce cas, les intersections avec certaines surfaces coordonnées (ici les plans $z = h$) sont alors des cercles. Si tous les demi-axes sont égaux, on obtient évidemment une sphère. Un ellipsoïde est la surface décrite par une infinité d'ellipses dont les sommets se déplacent sur deux autres ellipses.



Hyperboloïde à une nappe

Son équation, lorsque sa position par rapport aux axes de coordonnées est celle de la *figure* 16, est :

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}-\frac{z^2}{c^2}=1$$

et les sections principales :

plan $z=0$: $\frac{x^2}{a^2}+\frac{y^2}{b^2}=1$ (ellipse de demi-axes a et b)

plan $y=0$: $\frac{x^2}{a^2}-\frac{z^2}{c^2}=1$ $\left(\text{hyperbole d'asymptotes } z=\pm\frac{c}{a}x\right)$

plan $x=0$: $\frac{y^2}{b^2}-\frac{z^2}{c^2}=1$ $\left(\text{hyperbole d'asymptotes } z=\pm\frac{c}{b}y\right)$

Cet hyperboloïde admet les mêmes éléments de symétrie que l'ellipsoïde, son intersection avec les axes Ox et Oy (il ne coupe pas Oz) nous donne les quatre sommets $(\pm a, 0, 0)$ et $(0, \pm b, 0)$. L'hyperboloïde à une nappe est la surface décrite par des ellipses dont les sommets se déplacent sur deux hyperboles, deux sommets opposés se déplaçant chacun sur une branche de la même hyperbole. On peut démontrer à partir de l'équation qu'il s'agit d'une surface réglée : par chaque point de sa surface passent deux droites distinctes lui appartenant tout entières. Il est engendré par deux familles de droites telles que deux droites prises chacune dans une famille se rencontrent toujours alors que deux droites de la même famille ne se rencontrent jamais. Si $a=b$, les sections par les surfaces coordonnées $z=h$ sont des cercles, et l'hyperboloïde est alors de révolution autour de Oz.

Hyperboloïde à deux nappes

C'est la surface représentée sur la *figure* 17.

Avec un système de coordonnées bien choisi, son équation est :

$$-\frac{x^2}{a^2}+\frac{y^2}{b^2}-\frac{z^2}{c^2}=1$$

Il admet comme sections principales :

$$\text{plan } z=0: -\frac{x^2}{a^2}+\frac{y^2}{b^2}=1$$

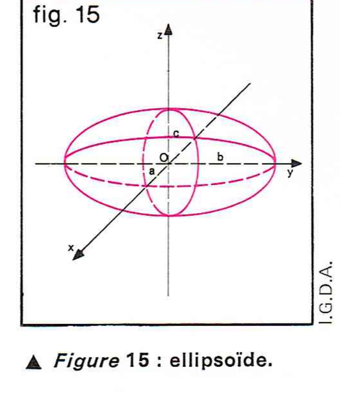
$$\left(\text{hyperbole d'asymptotes } y=\pm\frac{b}{a}x\right)$$

$$\text{plan } x=0: \frac{y^2}{b^2}-\frac{z^2}{c^2}=1$$

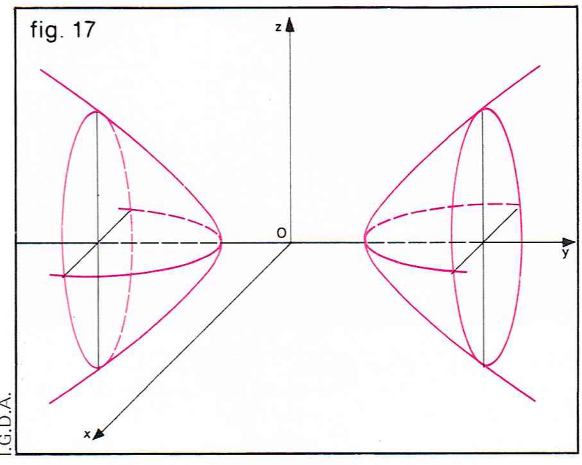
$$\left(\text{hyperbole d'asymptotes } z=\pm\frac{c}{b}y\right)$$

Le plan $y=0$ ne rencontre pas la surface ; les sections par des plans $y=k$ avec $k<-b$ ou $k>b$ sont des ellipses. Ainsi, cet hyperboloïde est la surface décrite par des ellipses dont les sommets se déplacent sur deux hyperboles perpendiculaires, mais cette fois deux sommets opposés se déplacent sur la même branche d'hyperbole.

Les éléments de symétrie sont les mêmes que dans le cas de l'ellipsoïde et de l'hyperboloïde à une nappe.



▲ ***Figure* 15 : ellipsoïde.**



◀ ***A gauche, figure* 16 : hyperboloïde à une nappe. *A droite, figure* 17 : hyperboloïde à deux nappes.**

Paraboloïde elliptique

Si sa disposition par rapport aux axes est celle de la *figure* 18, son équation est :

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=2z$$

Il admet les plans xOz et yOz comme plans de symétrie, l'axe Oz comme axe de symétrie. Les sections principales sont :

$$\text{plan } x=0: \frac{y^2}{b^2}-2z=0$$

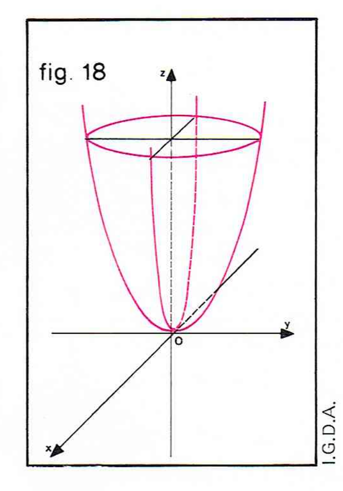
(parabole d'axe Oz de sommet l'origine)

$$\text{plan } y=0: \frac{x^2}{a^2}-2z=0$$

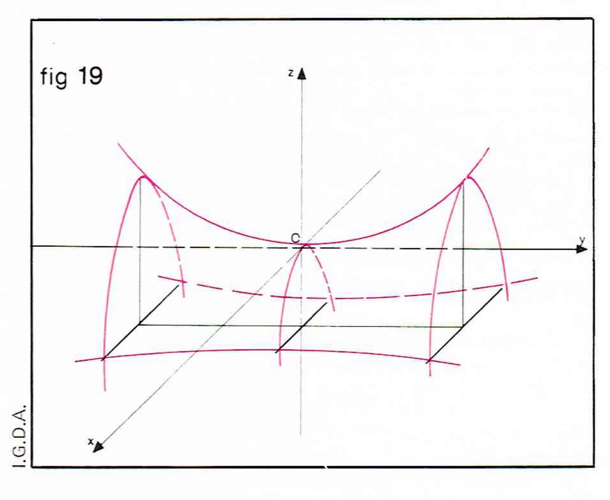
(parabole d'axe Oz de sommet l'origine)

L'intersection avec le plan $z=0$ est réduite à l'origine, les sections par des plans $z=k>0$ sont des ellipses centrées sur l'axe Oz.

Le paraboloïde elliptique est la surface décrite par des ellipses dont les sommets se déplacent sur deux paraboles perpendiculaires, de même sommet, de même axe, situées du même côté du plan $z=0$.

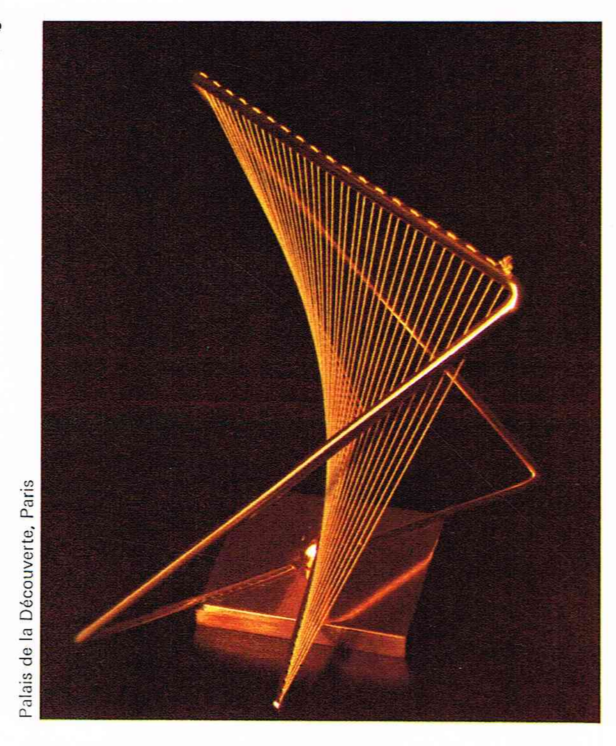


▲ ***Figure* 18 : paraboloïde elliptique.**



◀ ***Figure* 19 : paraboloïde hyperbolique.**

▶ *Un paraboloïde hyperbolique.*


Palais de la Découverte, Paris

Paraboloïde hyperbolique

Son équation, dans le système d'axes de la *figure* 19, s'écrit :

$$-\frac{x^2}{a^2}+\frac{y^2}{b^2}=2z$$

Il admet les mêmes éléments de symétrie que le paraboloïde elliptique; les sections principales sont :

plan $x=0: \frac{y^2}{b^2}-2z=0$

(parabole d'axe Oz de sommet l'origine située du côté des z positifs)

plan $y=0: -\frac{x^2}{a^2}-2z=0$

(parabole d'axe Oz de sommet l'origine située du côté des z négatifs)

plan $z=0: y=\pm\frac{b}{a}x$ (deux droites)

plan $z=k^2: -\frac{x^2}{a^2}+\frac{y^2}{b^2}=2k^2$

(hyperbole d'axe focal parallèle à Oz)

plans $z=-k^2: \frac{x^2}{a^2}-\frac{y^2}{b^2}=2k^2$

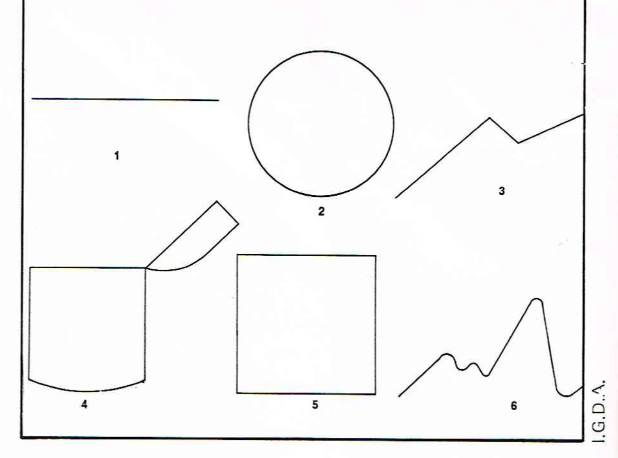
(hyperbole d'axe focal parallèle à Ox)

Un paraboloïde hyperbolique est la surface obtenue lorsqu'on déplace une parabole dans une direction perpendiculaire à son plan, son sommet décrivant une autre parabole, de concavité opposée, située dans un plan perpendiculaire à celui de la première et d'axe parallèle à celui de celle-ci.

On a vu qu'il passait par le point O deux droites appartenant à la surface, et on peut démontrer que c'est une propriété que possèdent tous les points du paraboloïde hyperbolique. De même que l'hyperboloïde à une nappe, il est engendré par deux familles de droites ayant les mêmes caractéristiques.

BIBLIOGRAPHIE

CASANOVA, *Cours de mathématiques spéciales,* t. III, *Géométrie analytique,* Belin, 1965. - DELACHET, *la Géométrie analytique,* « Que sais-je? », n° 1047, P. U. F., 1967. - QUINET, *Cours élémentaire de mathématiques supérieures,* t. VI, *Géométrie analytique plane,* Dunod, 1966.



I.G.D.A.

TOPOLOGIE

Considérons les six courbes de la *figure* 1 ; elles ne sont pas équivalentes d'après les critères de la géométrie élémentaire. Toutefois, on remarque intuitivement que les courbes 1, 3 et 6 ont des propriétés communes que les autres ne possèdent pas : on peut les déformer continûment l'une dans l'autre, mais aucune d'entre elles ne peut être déformée en les autres. On peut en dire de même des courbes 2 et 5.

On dit que les courbes 1, 3 et 6 sont *topologiquement équivalentes*, ainsi que les courbes 2 et 5; tandis que la courbe 4 ne l'est à aucune des autres. Par analogie, les surfaces délimitées par un cube ou une sphère sont topologiquement équivalentes. La topologie est la partie des mathématiques qui étudie cette notion intuitive de continuité et de limite.

Jusqu'au début du XIXe siècle, les mathématiciens utilisèrent les notions de limite et de continuité sans les définir rigoureusement. C'est à cette époque qu'*A. Cauchy, N. Abel* et *B. Bolzano* définirent la limite d'une suite numérique et la continuité d'une fonction numérique d'une variable numérique; ce fut le début de la topologie.

G. Hilbert chercha à axiomatiser les notions de limite et de continuité; il introduisit pour cela les voisinages. Mais ce furent M. Fréchet et F. Riesz qui, au début du XXe siècle, firent les premières études axiomatiques des notions de point limite et de continuité. Tandis que M. Fréchet fondait son étude sur une notion générale de distance, F. Riesz procédait à une définition axiomatique directe de la notion de point limite et, de cette façon, arrivait le premier à la notion d'espace topologique. Un peu plus tard, F. Hausdorff donna aux axiomes de la topologie une forme à peu près identique à celle qui est utilisée aujourd'hui. La notion de compact fut dégagée par P. S. Alexandroff, P. Urysohn et A. Tychonov. Enfin, vers 1940, la définition des filtres, par H. Cartan, mettait un point final à l'histoire de la notion de limite.

Les espaces métriques et les espaces topologiques

Les espaces métriques

La notion d'espace métrique, introduite par *M. Fréchet* en 1906 et développée peu après par *F. Hausdorff,* est directement issue d'une analyse de la distance usuelle.

L'analyse des principales propriétés de la distance entre deux points dans l'espace euclidien conduit à la définition axiomatique suivante : on appelle distance sur un ensemble E une application d de E $\times$ E dans l'ensemble $\mathbb{R}_+$ des nombres réels positifs ou nuls tels que, quels que soient les éléments x, y et z de E, on ait :

a) $d(x, y) = 0 \Leftrightarrow x = y;$

b) $d(x, y) = d(x, y);$

c) $d(x, y) \leqslant d(x, z) + d(z, y).$

Cette dernière propriété est appelée *inégalité triangulaire*, car elle est la généralisation de la classique inégalité entre les longueurs des côtés d'un triangle : tout côté d'un triangle est au plus égal à la somme des deux autres.

Un ensemble E muni d'une telle distance est appelé *espace métrique.*

Remarquons que $d(x, y)$ est bien un nombre réel positif ou nul : d'après **a)**, **c)** et **b)**, on a en effet pour tout x et y appartenant à E :

$$0 = d(x, x) \leqslant d(x, y) + d(y, x) = 2\,d(x, y)$$

On démontre aussi l'inégalité :

$$d(x, z) \geqslant |d(x, y) - d(y, z)|$$

(tout côté d'un triangle est au moins égal à la différence des deux autres).

Si (E, d) et (E', d') sont deux espaces métriques, une bijection f de E sur E' sera appelée *isométrie* si elle conserve la distance, c'est-à-dire si $d'[f(x), f(y)] = d(x, y)$ pour tout x et $y \in E$. Deux espaces métriques sont dits *isométriques* s'il existe une isométrie de l'un sur l'autre. Ils présentent alors « par transport » des propriétés semblables.

Donnons quelques exemples importants d'espaces métriques :

I) Soit $x = (x_1, x_2, ..., x_n)$ et $y = (y_1, y_2, ..., y_n)$ deux éléments de R^n, on pose :

$$d_2(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

On peut facilement vérifier que l'application $d_2 : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$ définie par $(x, y) \to d_2(x, y)$ est une distance sur R^n.

Par définition, d_2 est la distance euclidienne sur R^n.

On appelle aussi cette métrique la métrique naturelle de R^n.

II) Les applications m et s de $\mathbb{R}^n \times \mathbb{R}^n$ dans $\mathbb{R}_+$ définies par :

$$m(x, y) = \max_{1 \leqslant i \leqslant n} (|x_i - y_i|) \text{ et } s(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

sont aussi des distances sur $\mathbb{R}^n$. On démontre aisément que $m \leqslant d_2 \leqslant s \leqslant nm$. On en verra les conséquences plus loin.

III) Soit E un ensemble quelconque. L'application d de $E \times E \to \mathbb{R}$ définie ainsi :

$$d(x, y) = \begin{cases} 1 & \text{si } x \neq y \\ 0 & \text{si } x = y \end{cases}$$

est une distance sur E. On l'appelle *métrique discrète* ou *distance triviale*.

IV) Si E est un espace métrique muni d'une distance d, tout sous-ensemble A de E est un espace métrique, dit *sous-espace métrique* de E pour la distance induite d' définie ainsi :

$$d'(x, y) = d(x, y) \quad x, y \in A$$

V) Soit E un ensemble non vide et f une application injective de E dans $\mathbb{R}$. On peut montrer que l'application de $E \times E$ dans $\mathbb{R}$ définie par :

$$(x, y) \to d(x, y) = |f(x) - f(y)|$$

est une distance sur E.

On verra plus loin que les espaces métriques dont les éléments sont des fonctions sont parmi les plus importants (voir *Analyse fonctionnelle*).

VI) Une classe très importante d'espaces métriques est constituée par les *espaces vectoriels normés* en définissant la distance de deux éléments x et y comme la norme de leur différence, soit :

$$d(x, y) = \|x - y\|.$$

Rappelons la définition d'un *espace vectoriel normé* réel (resp. complexe) : soit E un espace vectoriel sur $\mathbb{R}$ (resp. $\mathbb{C}$) ; une *norme* sur E est une application de E dans $\mathbb{R}$, notée $x \to \|x\|$, telle que, quels que soient $x \in E$, $y \in E$, $\lambda \in \mathbb{R}$ (resp. $\mathbb{C}$), on ait les propriétés :

a) $\|x\| \geqslant 0, \quad \|x\| = 0 \Rightarrow x = 0$

b) $\|\lambda x\| = |\lambda| \, \|x\|$;

c) $\|x + y\| \leqslant \|x\| + \|y\|$ (inégalité de Minkowski).

Un espace vectoriel normé réel (resp. complexe) est la donnée d'un espace vectoriel E sur $\mathbb{R}$ (resp. sur $\mathbb{C}$) et d'une norme sur E. Ces espaces sont les espaces métriques dont les propriétés « ressemblent le plus » à celles des espaces numériques habituels.

VII) *La droite numérique achevée*

Désignons par $\overline{\mathbb{R}}$ la droite numérique achevée :

$$\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty\} \cup \{+\infty\},$$

qui est obtenue en adjoignant à l'ensemble $\mathbb{R}$ des nombres réels deux nouveaux éléments que l'on désigne traditionnellement par $-\infty$ et $+\infty$, vu le rôle qu'ils jouent en analyse. Remarquons que l'application f définie par :

$$f(x) = \frac{x}{1 + x} \quad x \in \mathbb{R},$$

$$f(+\infty) = +1 \text{ et } f(-\infty) = -1$$

est une bijection de $\overline{\mathbb{R}}$ sur le segment $[-1, +1]$. On peut donc transporter la distance usuelle sur $\mathbb{R}$, en définissant une distance sur $\overline{\mathbb{R}}$ par : $d'(x, y) = |f(x) - f(y)|$; et bien entendu, f est une isométrie de $\overline{\mathbb{R}}$, muni de cette distance d', sur le segment fermé $[-1, +1]$ muni de la distance habituelle.

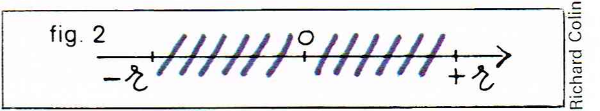
On verra que, dans un espace métrique, on peut traiter d'un grand nombre de notions de l'analyse mathématique. Pour cela, on a besoin de donner quelques définitions.

On appelle *boule ouverte* (resp. *fermée*) de centre a et de rayon r fini > 0, dans un espace métrique (E, d), et on note $B(a, r)$ (resp. $\overline{B}(a, r)$) l'ensemble des points E dont la distance à a est inférieure (resp. inférieure ou égale) à r.

Exemples

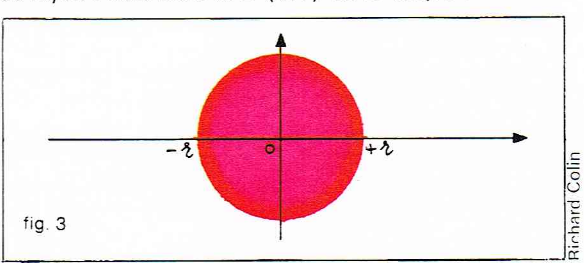
I) Considérons dans $\mathbb{R}^n$ muni de la distance euclidienne l'ensemble : $B(O, r) = \{x \in \mathbb{R}^n : d_2(O, x) < r\}$ (*fig.* 2, 3).

Pour $n = 1$: $B(O, r) = \{x \in \mathbb{R} : |x| < r\}$, la boule ouverte de centre O et de rayon r n'est autre que l'intervalle ouvert $]-r, +r[$ et on a : $\overline{B}(O, r) = [-r + r]$.



◀ **Figure 2.**

Pour $n = 2$: $B(O, r)$ est le disque de centre O et de rayon r sans bord et $\overline{B}(O, r)$ est le disque avec bord.



◀ **Figure 3.**

Pour $n = 3$: nous rencontrons avec $B(O, r)$ une boule au sens usuel.

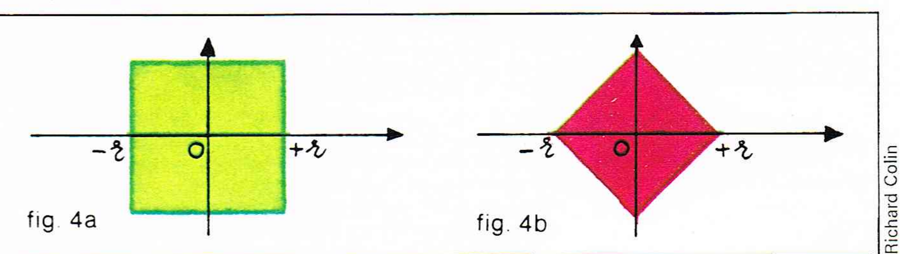
II) Considérons maintenant les boules de centre O et de rayon r dans $\mathbb{R}^2$ muni des distances m et s définies précédemment :

$B_m(O, r)$ n'est autre qu'un carré de centre O (*fig.* 4 a).

$B_s(O, r)$ est le carré ci-dessous (*fig.* 4 b) :



▼ ***Figures* 4a *et* 4b.**



III) Dans un ensemble E muni de la métrique discrète, une boule fermée de rayon < 1 se réduit à son centre, une boule fermée de rayon $\geqslant 1$ est l'espace tout entier.

Une partie d'un espace métrique est dite *bornée* si elle est contenue dans au moins une boule de rayon fini : ainsi $\mathbb{R}$ n'est pas borné ; tout ensemble muni de la métrique discrète est borné.

On dit que deux distances d et d' définies sur un même espace métrique sont *équivalentes* s'il existe deux constantes $a, b \in \mathbb{R}$ telles que :

$$d(x, y) \leqslant ad'(x, y) \quad \text{et} \quad d'(x, y) \leqslant bd(x, y)$$

pour tous $x, y \in E$.

On peut constater que les distances d_2, m et s définies sur $\mathbb{R}^n$ sont équivalentes puisque :

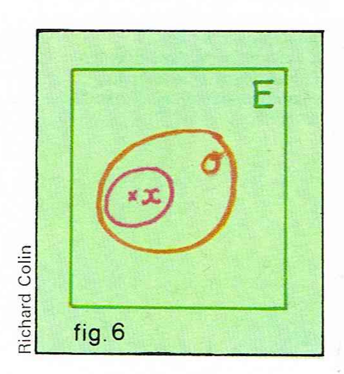
$$m \leqslant d_2 \leqslant s \leqslant nm.$$

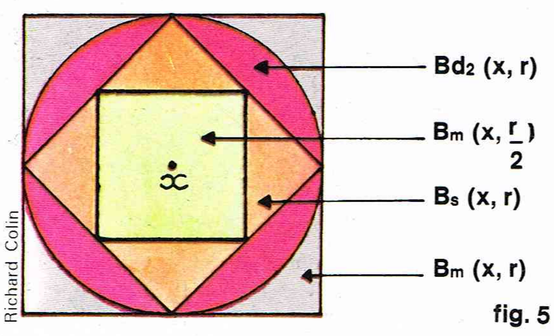
◀ ***Page ci-contre, à droite ; figure* 1 *: les courbes* 1, 3, 6 *sont* topologiquement équivalentes *parce qu'on peut les déformer* continûment *l'une dans l'autre ; de même pour les courbes* 2 *et* 5.**
La courbe* 4, *par contre, n'est topologiquement équivalente à aucune des autres.

Ces inégalités se traduisent par les inclusions :

$$B_m\left(x, \frac{r}{n}\right) \subset B_s(x, r) \subset B_{d2}(x, r) \subset B_m(x, r).$$

On a pour $n = 2$ (*fig.* 5) :





▲ *A gauche, figure* **6** *; à droite, figure* **5.**

Soit (E, d) un espace métrique et $\mathcal{O}$ une partie de E, on dit que $\mathcal{O}$ est un *ouvert* si, quel que soit $x \in \mathcal{O}$, il existe une boule ouverte de centre x contenu dans $\mathcal{O}$ (*fig.* 6).

Toute boule ouverte B (a, r) d'un espace métrique est un ouvert : soit en effet un élément x d'une telle boule, le nombre $r_x = r - d(a, x)$ est positif ; et $y \in B(x, r_x)$, on a :

$$d(y, a) \leqslant d(y, x) + d(x, a) < r_x + d(a, x) = r,$$
$$\text{donc } B(x, r_x) \subset B(a, r)$$

D'après ce qu'on vient de voir, un sous-ensemble de E ouvert quand E est muni d'une certaine distance d, l'est aussi quand E est muni d'une distance d' équivalente à d.

Soit A une partie d'un espace métrique (E, d), on dit que A est *fermée* dans E si son complémentaire dans E est un ouvert et on appelle *fermeture* (ou *adhérence*) de A et on note $\overline{A}$ le plus petit fermé contenant A. Si $x \in \overline{A}$, on dit que x est *adhérent* à A. On dit que A est *dense* dans E si $\overline{A} = E$.

On appelle *intérieur* de A et on note $\mathring{A}$ le plus grand ouvert contenu dans A. Si $x \in \mathring{A}$, on dit que x est *intérieur* à A.

La *frontière* de A (notée Fr A) est l'ensemble $\overline{A} \cap \overline{\complement_E A}$.

Un espace métrique E est dit *séparable* s'il est fini ou s'il contient une partie dénombrable dense dans lui-même.

Les espaces topologiques

Appelons topologie sur un ensemble muni d'une métrique l'ensemble de ses parties ouvertes ; alors, deux métriques distinctes sur un même ensemble peuvent ainsi donner lieu à la même topologie (on a vu que c'était le cas lorsqu'elles sont équivalentes). Beaucoup de notions ou de définitions ne dépendent pas de la métrique choisie, mais seulement de la topologie ; par exemple, les notions de fermé, de fermeture, d'intérieur, de frontière et la notion de limite (qui sera définie plus loin) ; on dit que ce sont des *notions topologiques*. Par contre, les notions de boules, d'ensemble borné, de suites de Cauchy, d'espace complet (ces deux dernières notions seront définies plus loin) dépendent essentiellement de la distance choisie ; elles peuvent être altérées si on passe à une distance équivalente ; on les appelle des *notions métriques*.

Les notions et propriétés topologiques ne dépendent que des trois propriétés suivantes de la famille des ouverts :

(O1) E et ∅ sont des ensembles ouverts ;

(O2) toute réunion *(finie ou infinie)* d'ensembles ouverts est un ensemble ouvert ;

(O3) toute intersection *finie* d'ouverts est un ouvert.

Ceci suggère de les prendre comme *axiomes* d'une nouvelle théorie, celle des *espaces topologiques*, théorie généralisant et englobant celle des propriétés topologiques des espaces métriques.

On appelle *espace topologique* E la donnée d'un ensemble E et d'une famille de parties, appelées parties ouvertes de la topologie, vérifiant les trois propriétés **(O1)**, **(O2)** et **(O3)**.

Remarque : par passage au complémentaire, il résulte immédiatement des énoncés **(O1)**, **(O2)** et **(O3)** trois énoncés **(F1)**, **(F2)** et **(F3)** équivalents aux précédents et concernant les fermés de E :

(F1) ∅ et E sont des ensembles fermés ;

(F2) toute intersection *(finie ou infinie)* de fermés est fermée ;

F(3) toute réunion *finie* de fermés est fermée.

Exemples

— Sur un ensemble quelconque E, prenons pour parties ouvertes l'ensemble de toutes les parties de E. Cette topologie est dite *discrète*. Prenons pour parties ouvertes l'ensemble des deux seules parties ∅ et E : on obtient la topologie *grossière*.

— Tout espace métrique définit évidemment un espace topologique, si on prend comme ouverts les parties que nous avons appelées ouvertes dans l'espace métrique. Cette topologie est dite *topologie naturelle*.

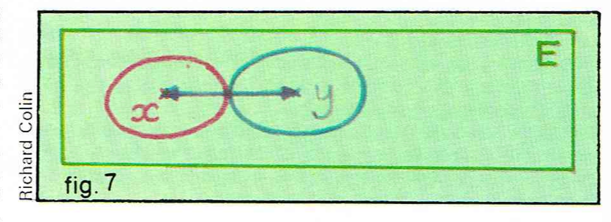
— Soit E un espace topologique et F une partie de E ; on peut faire de F un espace topologique, en prenant comme ouverts les intersections avec F des ouverts de la topologie de E. F est alors appelé *sous-espace topologique* de E, et sa topologie est dite topologie *induite* par celle de E.

— Si E et F sont deux espaces topologiques, on appelle pavé ouvert de $E \times F$ tout sous-ensemble de la forme $U \times V$ où U est un ouvert de E et V un ouvert de F. On vérifie facilement que ces ensembles ainsi définis sur le produit possèdent tous les axiomes que doivent vérifier les ensembles ouverts d'une topologie. Cette topologie s'appelle *topologie produit*. Remarquons que la topologie naturelle de $\mathbb{R}^2$ est la topologie naturelle de $\mathbb{R}$ par elle-même.

— Soit E un espace topologique et ρ une relation d'équivalence sur E. Dans l'ensemble quotient E/ρ, on appelle ouverts les ensembles de classes d'équivalence tels que leur réunion soit un ouvert de E. Cette topologie s'appelle *topologie quotient*. En particulier, si E est $\mathbb{R}^{n+1} - \{0\}$, c'est-à-dire $\mathbb{R}^{n+1}$ auquel on supprime le point O de coordonnées toutes nulles, et si ρ est la relation qui associe les points de coordonnées proportionnelles, on définit la topologie de l'espace projectif réel à n dimensions $P_n(\mathbb{R})$ (voir *Géométrie*).

Un espace topologique est dit *séparé* si, quels que soient les points x et y distincts de E, il existe deux ouverts d'intersection vide, l'un contenant x et l'autre y (propriété de séparation de Hausdorff).

On remarquera que tout espace métrique est séparé : en effet, soit les deux boules ouvertes de centre x et y et de rayon $\frac{d(x, y)}{2}$; on démontre facilement qu'elles sont d'intersection vide (*fig.* 7).



▶ *Figure* **7.**

Cette propriété est très intéressante, car elle permet — comme on le verra plus loin — de montrer l'unicité de la limite d'une suite dans un tel espace. *Les espaces topologiques non séparés n'ont qu'un usage très limité en analyse*.

On dit qu'un espace topologique est *métrisable* s'il existe une métrique qui engendre sa topologie ; mais une telle métrique n'est pas donnée, car s'il en existe une, il en existe une infinité (sauf si $E = \varnothing$) : si une distance d définit sa topologie, kd $(k \in \mathbb{R}_+)$ la définit aussi.

Un espace topologique métrisable est dit *régulier*. On peut vérifier que la topologie discrète est métrisable, une des métriques étant la métrique discrète ; par contre, la topologie grossière n'est pas métrisable, car elle n'est même pas séparée.

Soit E un ensemble dans lequel sont définies deux topologies T_1 et T_2. On dit que T_1 est plus fine ou plus forte que T_2 si toute partie ouverte pour T_2 est aussi ouverte pour T_1 (c'est-à-dire que l'ensemble des ouverts de T_1 contient l'ensemble des ouverts de T_2). La topologie discrète est, par exemple, plus fine que toutes les autres topologies sur E ; la topologie grossière est la moins fine de toutes les topologies.

Voisinages et continuité

En analyse, une des premières approches de la notion de continuité repose sur l'idée qu'une fonction continue f d'une variable réelle x doit être telle qu'à de petites variations de x correspondent de petites variations de $f(x)$; cette condition exprime en fait que le graphe $y = f(x)$ doit être, au sens intuitif, une courbe continue. Si on conserve un point de vue géométrique, en notant d'abord que la fonction f peut être interprétée comme une application de l'axe des x sur l'axe des y, chaque valeur de x étant appliquée sur une valeur $f(x)$ de l'axe y déterminée de manière unique. On peut alors exprimer la continuité de f en x' en disant que des points *suffisamment proches* de x' sur l'axe des x sont appliqués par f en des points *arbitrairement proches* $f(x')$ sur l'axe des y.

Introduisons maintenant les *voisinages* pour préciser la notion de continuité : on appelle *voisinage* d'un point a d'un espace topologique E toute partie de E contenant au moins un ouvert contenant lui-même a.

Soit $(V_i)_{i \in I}$ une famille de voisinages de a dans E. On dit que c'est un *système fondamental de voisinages* de a si tout voisinage de a contient l'un des V_i. On démontre le théorème suivant : un espace topologique E est séparé si, et seulement si, pour tout $a \in E$, l'intersection de tous les voisinages fermés de a se réduit à $\{a\}$ (*fig.* 8).

De par sa définition, on voit que la notion de voisinage de a est très proche de celle d'ouvert contenant a, avec laquelle elle fait souvent double emploi.

Cette nouvelle notion de voisinage permet d'introduire celles de *limite* et de *continuité*.

Soit E et E′ deux espaces topologiques, A une partie de E et a un point de E adhérent à A ; on dit, si f est une application de A dans E′, que $f(x)$ *tend vers la limite* l *lorsque* x *tend vers* a *par valeurs dans A* si, quel que soit le voisinage V′ de l dans E′, il existe un voisinage V de a dans E tel que :

$$f(V \cap A) \subset V'.$$

Exemples

— Ainsi, si f est une application $\mathbb{R}$ dans E′, l'expression $f(x)$ tend vers l lorsque x tend vers a par valeur supérieure signifie que : quel que soit le voisinage V′ de l dans E′, il existe $\eta > 0$ tel que $(|x - a| \leqslant \eta, x \geqslant a)$ entraîne $f(x) \in V'$. Ici $E = \mathbb{R}$ et $A = [a, +\infty[$.

— Dans les mêmes conditions, l'expression $f(x)$ tend vers l lorsque x tend vers $+\infty$, signifie que, quel que soit le voisinage V′ de l dans E, il existe α réel tel que $x \geqslant \alpha$ entraîne $f(x) \in V'$. Ici $E = \overline{\mathbb{R}}$, $A = \mathbb{R}$, $a = +\infty$.

Soit f une application d'un espace topologique E dans un espace topologique E′. On dit que f est *continue* en un point a de E si, quel que soit V′ voisinage de $f(a)$, il existe V, voisinage de a, tel que : $f(V) \subset V'$;

ou encore :

si l'image réciproque par f de tout voisinage de $f(a)$ est un voisinage de a ;

ou encore :

si $f(x)$ tend vers $f(a)$ quand x tend vers a.

S'il s'agit d'espaces métriques, on peut encore dire :

si, quel que soit $\varepsilon > 0$, il existe $\eta > 0$ tel que $d(a, x) < \eta$ entraîne $d(f(a), f(x)) < \varepsilon$;

ce qui est encore équivalent à :

si, quel que soit la boule de centre $f(a)$, il existe une boule de centre a dont l'image par f appartienne à la précédente (*fig.* 9).

On démontre aisément que l'application composée de deux applications continues est continue.

Cette notion de continuité est *locale*, car elle ne dépend que des voisinages du point considéré. On remarquera qu'elle est une généralisation de la continuité telle qu'elle était connue pour $E = E' = \mathbb{R}$.

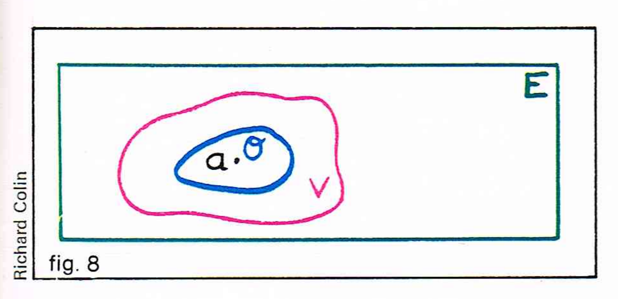
Une application de E dans E′ est dite *continue* si elle est continue en tout point de E.

Si une application d'un espace topologique E dans un espace topologique E′ est continue en un point a de E, elle le reste *a fortiori* si on remplace la topologie de E par une topologie plus fine, et celle de E′ par une moins fine.

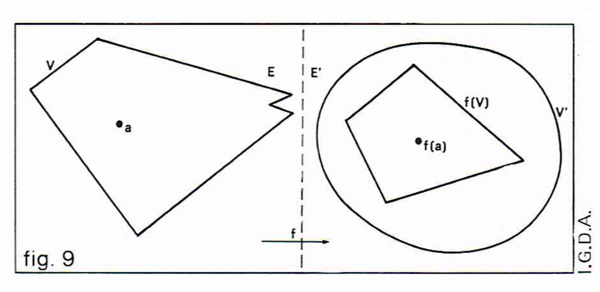
Remarquons que, si (E, d) et (E', d') sont des espaces métriques et si une application f de E dans E′ est continue en un point a de E, elle le reste si on remplace d et d' par des distances équivalentes. Notons aussi que f est

M.C.Escher, *le Ruban de Möbius* II - 1963 - Escher Foundation - Haags Gemeentemuseum - The Hague

▲ ***Le ruban de Möbius, ruban à une seule face, dans une interprétation artistique de M. C. Escher.***

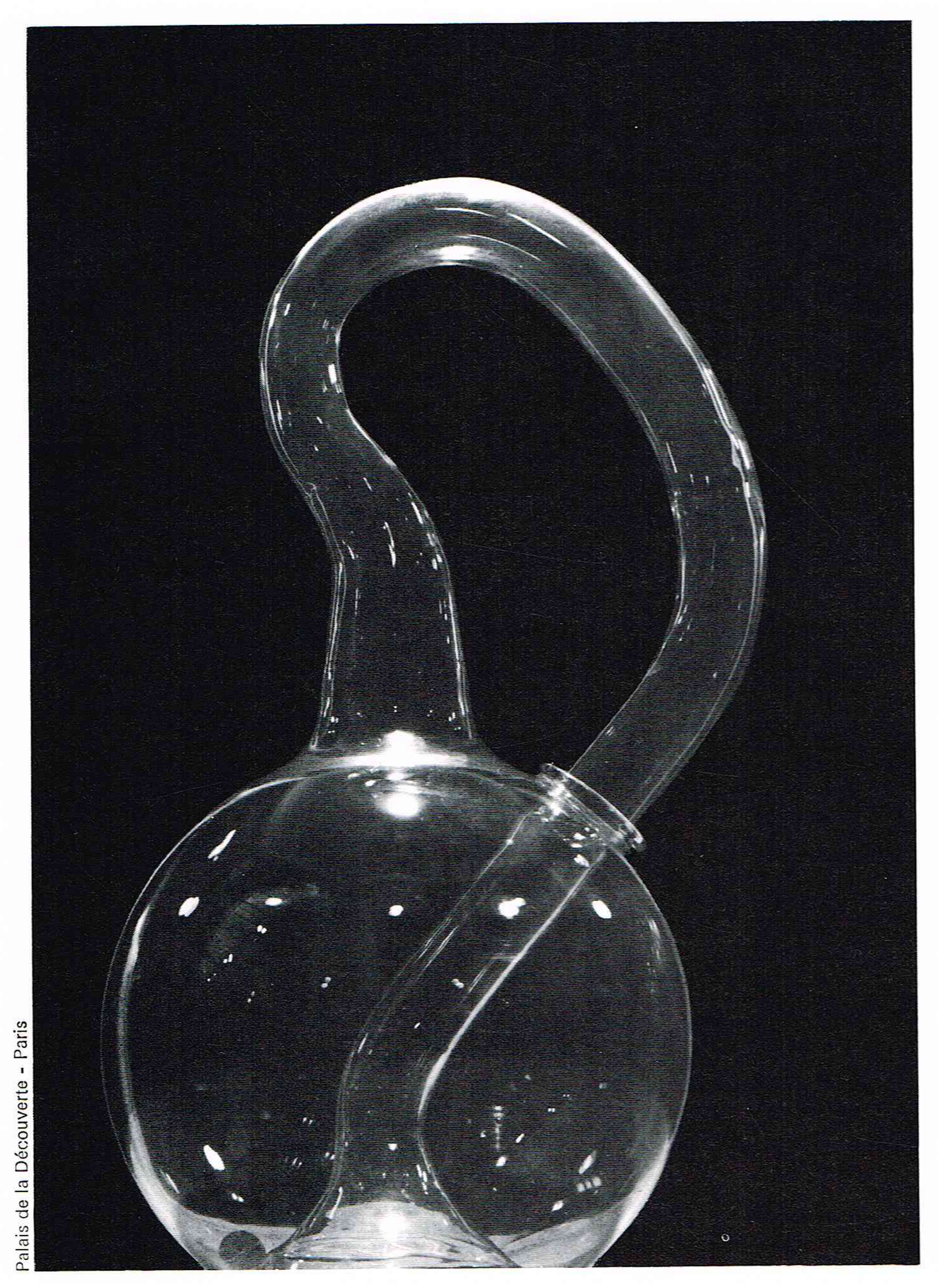


Richard Colin



I.G.D.A.

◄ ***A gauche, figure 8 : notion de voisinage. A droite, figure 9 : application continue au point a d'un espace E dans un espace E′.***



▲ ***La* bouteille de Klein *imaginée par le mathématicien allemand Félix Klein en 1882 : cette curieuse bouteille, dont le col traverse la paroi et vient faire jonction avec le fond, ne comporte ni face interne ni face externe.***

toujours continue si E est un espace topologique discret, ou si E′ est un espace topologique grossier.

Dans tout espace métrique (E, d), la fonction distance en un point donné a : $x \to d(a, x)$, application de E dans $\mathbb{R}_+$, est partout continue ; cela résulte de ce que :

$$|d(a, x) - d(a, x_0)| \leqslant d(x_0, x).$$

On a un théorème important pour les applications continues d'un espace topologique E dans un espace topologique E′ : l'image réciproque par f de tout ouvert de E′ est un ouvert de E et l'image réciproque de tout fermé de E′ est un fermé de E. Signalons que, par une application continue, l'image directe d'un ouvert n'est pas ouverte en général.

Étudions maintenant des applications particulièrement importantes en topologie : les homéomorphismes, dont on a déjà signalé l'intérêt au début de cette partie.

On appelle **homéomorphisme** d'un espace topologique E sur un espace topologique E′ toute bijection de E sur E′ telle qu'elle-même et sa réciproque soient continues.

Remarquons qu'il ne faut pas croire que toute application bijective et continue soit nécessairement un homéomorphisme.

Exemple

Soit E = $[a, b[\cup \{c\}$ où a, b et $c \in \mathbb{R}$ et $c > b$, et soit f une application de E dans E′ = $[a, b]$ telle que :

$$f(t) = \begin{cases} t & \text{si } t \neq c \\ b & \text{si } t = c \end{cases}$$

Cette application est bijective et continue, mais son application réciproque n'est pas continue.

On dit que deux espaces topologiques E et E′ sont *homéomorphes* s'il existe au moins un homéomorphisme de l'un sur l'autre. *Ces deux espaces ont alors les mêmes propriétés topologiques*, c'est-à-dire *les mêmes propriétés pour tout ce qui concerne les ensembles ouverts, les ensembles fermés et les voisinages.*

Exemples

— L'intérieur d'un disque et l'intérieur d'un triangle dans un plan euclidien sont des espaces métriques homéomorphes.

— Le demi-plan $y > 0$, la région $y > x^2$ située au-dessus de la parabole $y = x^2$, la région $y < x^2$ située au-dessous de cette parabole, dans le plan $\mathbb{R}^2$, sont homéomorphes.

— La droite réelle, munie de sa métrique naturelle, et la droite réelle, munie de sa métrique discrète, ne sont pas homéomorphes, puisque, sur cette dernière, toutes les parties sont ouvertes et qu'il n'en est pas de même sur la première.

— On peut vérifier que l'application f de $\overline{\mathbb{R}}$ dans $[-1, +1]$ qui à $x \in \mathbb{R}$ fait correspondre $\frac{x}{1 + |x|}$, est un homéomorphisme, donc $\overline{\mathbb{R}}$ et $[-1, +1]$ sont homéomorphes.

Notons que les homéomorphismes d'un espace sur lui-même forment un groupe et, dans une classe d'espaces, la relation : « il existe un homéomorphisme entre E et E′ », est une relation d'équivalence.

Toutes les notions précédentes de continuité et d'homéomorphie sont topologiques. Il est utile d'avoir une notion de continuité plus forte, qui n'est plus seulement topologique, utilisant le fait qu'on peut « comparer » les voisinages de deux points distincts d'un espace métrique grâce aux rayons des boules. Plus précisément, on dira qu'une application f : (E, d) → (E′, d') est *uniformément continue* si :

quel que soit $\varepsilon > 0$, il existe $\eta > 0$, tel que :
$d(x, y) < \eta$ entraîne $d'(f(x), f(y)) < \varepsilon$;
ou encore :
quel que soit $\varepsilon > 0$, il existe $\eta > 0$, tel que :
$B(x, \eta) \subset f^{-1}(B(f(x), \varepsilon))$ quel que soit $x \in E$.

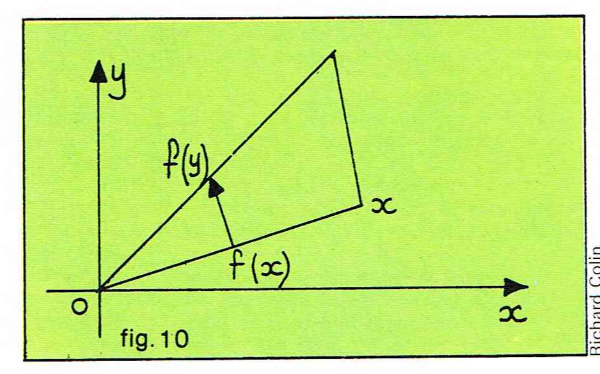
Par exemple, sur la droite réelle, la fonction x est uniformément continue, mais la fonction x^2 ne l'est pas ; en effet, on a :
$|(x + h)^2 - x^2| = |2hx + h^2| \geqslant 2hx$ si $h \geqslant 0, x \geqslant 0$;
alors pour ε donné, si on cherche h tel qu'on ait :

$$|2hx + h^2| \leqslant \varepsilon,$$

on doit nécessairement prendre $|h| \leqslant \frac{\varepsilon}{2|x|}$ et il est par suite impossible de choisir η indépendant de x.

De même $(x, y) \to xy$ n'est pas une application uniformément continue de $\mathbb{R}^2$ dans $\mathbb{R}$.

Un cas particulier de cette situation est fourni par les applications *lipschitziennes :* on dit qu'une application f : (E, d) → (E′, d') est lipschitzienne de rapport k ($k \in \mathbb{R}_+$) si on a :
$d'(f(x), f(y)) \leqslant k\,d(x, y)$ quels que soient $x, y \in E$;
ainsi, l'image réciproque par f de la boule de centre $f(x)$ et de rayon r contient la boule de centre x et de rayon r/k. Si $k = 1$, on a une *isométrie ;* si $k < 1$, on dit que l'application est *contractante* (*fig.* 10).



▶ ***Figure 10 :* *la fonction* f(x) = $\frac{x}{2}$ *est un exemple de* fonction contractante.**

Par exemple, l'application de $\mathbb{R}^2$ dans $\mathbb{R}^2$ qui à $x = (x_1, x_2)$ fait correspondre $\left(\frac{x_1}{2}, \frac{x_2}{2}\right)$ est contractante de rapport $\frac{1}{2}$.

Limites de suites — Espaces métriques complets

Soit E un espace topologique. Soit $x_0, x_1, \ldots, x_n, \ldots$, une *suite* de points de E. C'est une application de l'ensemble des entiers $\mathbb{N}$ dans E. On dit que la suite (x_n) *converge vers un point l de E* si, quel que soit le voisinage V de l, il existe un entier n_0 tel que, pour $n \geqslant n_0$, tous les x_n appartiennent à V.

Dans le cas où E est un espace métrique, on a encore : une suite (x_n) de (E, d) converge vers un point l de E ou a pour limite l si la suite des nombres réels : $d\ (l, x_0)$, $d\ (l, x_1)$, ..., $d\ (l, x_n)$, ..., converge vers 0 ;
ou encore :
quel que soit $\varepsilon > 0$, il existe un entier n_0 tel que, pour tout $n \geqslant n_0$, on ait : $d\ (l, x_n) \leqslant \varepsilon$.

Remarques

— Si une suite est convergente dans un espace topologique E, elle l'est *a fortiori* si on remplace la topologie de E par une topologie moins fine.

— Si une suite d'un espace topologique E admet une limite et si E est séparé, cette limite est nécessairement unique. On en déduit qu'étant donné qu'un espace métrique est toujours séparé, si une suite d'un tel espace admet une limite, cette limite est unique.

— On appelle suite extraite de la suite (x_n) toute suite de la forme $k \to x_{n_k}$, où $k \to n_k$ est une suite strictement croissante d'entiers $\geqslant 0$. *Si (x_n) converge vers x, toute suite extraite de (x_n) converge vers x*, la réciproque n'étant pas vraie.

— Pour qu'une application f d'un espace topologique E dans un espace topologique E′ soit continue en un point x de E, il faut que l'image par f de toute suite de points de E convergeant vers x soit une suite de points de E′ convergeant vers $f\ (x)$. Si E est métrisable, la condition est également suffisante.

— Pour qu'un point x d'un espace topologique E soit adhérent à une partie A de E, il suffit qu'il existe une suite d'éléments de A qui converge vers x. Si E est métrisable, cette condition est également nécessaire.

Ainsi, si E est métrisable, l'adhérence $\overline{A}$ d'une partie A de E est l'ensemble des limites des suites de A qui sont convergentes dans E. On en déduit immédiatement que, pour qu'une partie d'un espace topologique métrisable E soit fermée, il faut et il suffit qu'elle contienne toutes les limites de ses suites convergentes dans E.

Donnons une nouvelle définition importante : soit (E, d) un espace métrique. Une suite (x_n) dans E est appelée *suite de Cauchy* si :
quel que soit $\varepsilon > 0$, il existe $p \in \mathbb{N}$ tel que : pour tout $n \geqslant p$ et tout $m \geqslant p$, on ait : $d\ (x_n, x_m) < \varepsilon$;
ou encore :
$d\ (x_m, x_n)$ tend vers 0 quand n et m tendent vers $+\infty$.
On démontre facilement que toute suite de Cauchy est bornée.

On démontre aussi que, si une suite (x_n) a une limite x, alors (x_n) est de Cauchy. En effet, soit $p \in$ N tel que si $n \geqslant p$, on ait $d\ (x, x_n) < \frac{\varepsilon}{2}$; alors, si $n \geqslant p$ et $m \geqslant p$, on a :

$$d\ (x_n, x_m) \leqslant d\ (x_n, x) + d\ (x, x_m) < \varepsilon.$$

Mais, la réciproque de cette propriété n'est pas vraie en général, d'où la définition : un espace métrique E est dit *complet* si toute suite de Cauchy d'éléments de E a une limite dans E.

L'avantage de tels espaces est qu'on peut y montrer l'existence d'une limite sans connaître explicitement cette limite, procédé essentiel en analyse.

Exemples

— L'espace métrique E = [0, 1[pour la distance euclidienne n'est pas complet, car la suite $\left(x_n = 1 - \frac{1}{n}\right)$ est de Cauchy, mais n'a pas de limite dans E.

— Considérons la suite des entiers naturels : $x_n = n$; elle n'est manifestement pas de Cauchy pour la distance naturelle de $\mathbb{R}$, mais elle est une suite de Cauchy pour la d' déjà définie puisqu'elle converge vers $+\infty$ dans $\overline{\mathbb{R}}$:

$$d'\ (x, y) = \left|\frac{x}{1 + |x|} - \frac{y}{1 + |y|}\right|.$$

Un espace vectoriel normé complet est appelé *espace de Banach*. Ces espaces seront utilisés en analyse fonctionnelle.

Si E est un espace métrique et F un sous-espace métrique de E, alors : si F est complet, F est fermé dans E, et si E est complet et F fermé dans E, F est complet.

On démontre que $\mathbb{R}$ *muni de sa métrique naturelle est un espace métrique complet*. Ce résultat est très important en analyse et permet de déduire que :

— tout intervalle fermé de $\mathbb{R}$ est complet ;

— tout espace métrique produit de deux espaces complets étant complet, $\mathbb{R}^n$ *est complet* pour la métrique euclidienne ainsi que pour toutes les métriques qui lui sont équivalentes.

Espaces compacts

Si E est un espace topologique et $\{U_i, i \in I\}$ une famille de parties de E, on dit que cette famille recouvre E si

$$E = \bigcup_{i \in I} U_i$$

Un espace topologique E est dit *compact* s'il est séparé et si, de tout recouvrement de E par une famille d'ouverts $\mathcal{O}_i$, $i \in$ I, de E, on peut en extraire un recouvrement fini (c'est-à-dire qu'il existe une partie finie F de I telle que : $E = \bigcup_{i \in F} \mathcal{O}_i$).

On peut montrer par passage au complémentaire que cette dernière propriété est équivalente à la suivante : si une famille de fermés de E est d'intersection vide, alors il en existe une sous-famille dont l'intersection est déjà vide.

Remarques

— On en déduit immédiatement que, si E est un espace métrique compact, alors E est borné (il suffit de considérer un recouvrement de E par des boules ouvertes de même rayon).

— L'ensemble des intervalles ouverts $]n - 1, n + 1[$, $n \in$ Z, est un recouvrement ouvert de $\mathbb{R}$; il n'a pas d'autre sous-recouvrement que lui-même, car, si on supprime l'intervalle $]n - 1, n + 1[$, le point n n'est plus recouvert. Ceci prouve que $\mathbb{R}$ *n'est pas compact*.

— Si E est un espace topologique séparé et A une partie de E, alors : si A est compacte, A est fermée dans E, et si E est compact et A fermée dans E, A est compacte.

— Une réunion *fine* de parties compactes d'un espace topologique E séparé est une partie compacte de E.

Énonçons l'important **théorème de Borel-Lebesque :** *pour qu'une partie A de $\mathbb{R}$ soit compacte, il faut et il suffit qu'elle soit fermée et bornée dans $\mathbb{R}$.*

Conséquences

— Le même raisonnement où on remplace les intervalles par des hypercubes prouve que *les parties compactes de $\mathbb{R}^n$ sont exactement les parties fermées et bornées de $\mathbb{R}^n$.*

— On a vu que la droite achevée $\overline{\mathbb{R}}$ est homéomorphe à l'intervalle $[-1, +1]$. Il s'ensuit que $\overline{\mathbb{R}}$ *est compact, ainsi que $\overline{\mathbb{R}^n}$.*

Étudions maintenant l'importance de la notion de compacité pour les applications continues.

Si E et E′ sont deux espaces topologiques séparés et f une application continue de E dans E′, alors l'image f (K) de toute partie compacte K de E est compacte. Il s'ensuit que, si E est un espace compact, E′ un espace topologique séparé et f une application injective et continue de E dans E′, alors f est un homéomorphisme de E sur f (E).

On en déduit des propriétés intéressantes en analyse pour les applications continues à valeurs réelles.

— Si E est un espace compact et f une application continue de E dans $\mathbb{R}$, alors f est bornée et atteint ses bornes, c'est-à-dire qu'il existe $a \in$ E et $b \in$ E tels que, quel que soit $x \in$ E, on ait : $f\ (a) \leqslant f\ (x) \leqslant f\ (b)$;

En effet, f (E) est une partie fermée et bornée de $\mathbb{R}$. Les nombres réels $\sup_{x \in E} f\ (x)$ et $\inf_{x \in E} f\ (x)$ sont limites de suites d'éléments de f (E), donc appartiennent à f (E), car f (E) est fermé.



▶ *Le mathématicien français* **Émile Picard** *(1856-1941) à qui l'on doit, entre autres, une importante méthode de calcul des solutions d'équations par approximations successives : cette méthode est connue sous le nom de* **théorème du point fixe.**

— Si E est un espace compact et f une application continue de E dans $\mathbb{R}$ telle que $f(x) > 0$ pour tout $x \in$ E, alors il existe $m > 0$ tel que, pour tout $x \in$ E, on ait $f(x) \geqslant$ m. En effet, puisque $m = \inf_{x \in E} f(x) \in f(E)$, on a $m > 0$.

Si E est un espace compact et si (E', d') est un espace métrique, alors toute application continue de E dans E' est bornée. En effet, quel que soit $a \in$ E, $x \rightarrow d(a, f(x))$ est continue, donc bornée.

Théorème de Heine : *si (E, d) est un espace métrique compact, (E', d') un espace métrique et f une application continue de E dans E', alors f est uniformément continue.*

Soit une suite (x_n) d'éléments d'un espace topologique E ; on dit que a est *point adhérent à la suite* si, pour tout voisinage V de a, il existe une infinité de valeurs de l'entier n telles que $x_n \in$ V. Si une suite converge vers a, elle admet a comme point adhérent.

Énonçons une condition nécessaire et suffisante pour qu'un espace métrisable soit compact.

Théorème de Bolzano-Weierstrass : *si E est un espace métrisable, pour qu'il soit compact, il faut et il suffit que toute suite d'éléments de E admette au moins un point adhérent (ou que de toute suite, on puisse extraire une sous-suite convergente).*

Donnons sa conséquence immédiate : si E est un espace métrique compact, alors E est complet. Réciproquement, on montre qu'un espace métrique complet E est compact si et seulement s'il vérifie la propriété suivante de *précompacité :* pour tout $\varepsilon > 0$, il existe un recouvrement fini de E par des boules de rayon ε.

Un autre théorème important, qui sera utilisé en analyse fonctionnelle, est le **théorème de Tychonoff :** *tout produit, fini ou non, d'espaces compacts est compact. Inversement, si un produit d'espaces non vides est compact, chacun d'eux est compact.*

La démonstration de ce théorème repose sur la *théorie des filtres*, qui, bien que de plus en plus importante et utilisée, n'a pas été abordée dans cet exposé par souci de simplicité.

La notion de compacité est très forte ; il suffit souvent d'avoir à sa disposition des *espaces localement compacts* qui sont d'un usage constant en analyse : on dit qu'un espace topologique E est localement compact s'il est séparé et si chacun de ses points possède au moins un voisinage compact.

Remarques

— Tout espace métrique dont les boules fermées sont compactes est localement compact. Il s'ensuit que $\mathbb{R}^n$ *est localement compact*. Mais on doit bien noter qu'un espace métrique peut être localement compact (notion purement topologique) sans que ses boules fermées soient forcément compactes. On se convaincra de l'importance des espaces localement compacts en remarquant que tous les espaces de la géométrie : $\mathbb{R}$, $\mathbb{R}^n$, surfaces, courbes, sont localement compacts.

— Tout espace compact est évidemment localement compact ; tout espace topologique discret est localement compact.

— Si E est un espace localement compact et si A est une partie ouverte (resp. fermée) de E, alors le sous-espace topologique A de E est localement compact.

Application : **critères de non-homéomorphisme**

La compacité et la locale compacité étant invariantes par homéomorphismes, on peut montrer que deux espaces topologiques ne sont pas homéomorphes, parce qu'un de ces espaces est compact ou localement compact et que l'autre ne l'est pas. Par exemple, les intervalles [0, 1] et [0, 1[ne sont pas homéomorphes, parce que le premier est compact et que le second ne l'est pas. Un espace vectoriel de dimension finie est localement compact, par contre un espace vectoriel de dimension infinie ne l'est pas (voir le théorème de *F. Riesz* dans le chapitre *Analyse*) ; donc ils ne sont pas homéomorphes. Naturellement, ces critères sont très élémentaires, et la plupart des cas de non-homéomorphisme leur échappent. Par exemple, les deux espaces [0, 1[,]0, 1[sont tous deux localement compacts mais on démontre qu'ils ne sont pas homéomorphes. Deux espaces vectoriels normés de dimensions finies différentes sont tous deux localement compacts, mais ne sont pas homéomorphes (la démonstration est difficile).

La méthode des approximations successives

On doit à *É. Picard* une méthode de construction de solution d'équations par approximations successives formulée sous le nom de *théorème du point fixe*.

Pour le théorème des fonctions implicites et le théorème d'existence de la solution d'une équation différentielle, le théorème du point fixe est l'outil indispensable. Il est donc un des principaux théorèmes d'existence en mathématiques.

Théorème du point fixe : soit E un espace métrique *complet* et f une application contractante de E dans lui-même, c'est-à-dire qu'il existe une constante k, $0 < k < 1$, telle que : $d(f(x), f(y)) \leqslant kd(x, y)$ quels que soient x et y dans E. Alors l'équation $f(x) = x$ a une solution unique dans E. De plus, quel que soit $x_0 \in$ E, la suite (x_n)

définie par récurrence par $x_1 = f(x_0), \ldots, x_{n+1} = f(x_n), \ldots$ converge vers cette solution.

La démonstration est très simple, et on voit clairement le rôle joué par les suites de Cauchy. Remarquons d'abord que l'unicité est évidente : si $f(x) = x$ et $f(y) = y$, on a : $d(f(x), f(y)) = d(x, y) \leqslant kd(x, y)$, ce qui entraîne : $d(x, y) = 0$. Il suffit donc de montrer que (x_n) est convergente, car, si sa limite est x, on obtient $x = f(x)$ en faisant tendre n vers l'infini dans la relation de récurrence $x_{n+1} = f(x_n)$, puisque f est continue.

Or, on a :

$$d(x_{p+1}, x_p) = d(f(x_p), f(x_{p+1})) \leqslant kd(x_p, x_{p-1}),$$

d'où, par récurrence sur p :

$$d(x_{p+1}, x_p) \leqslant k^p d(x_1, x_0);$$

par inégalité triangulaire, on a pour $q \geqslant p$:

$$d(x_p, x_q) \leqslant \sum_{r=p}^{q-1} d(x_{r+1}, x_r) \leqslant \left(\sum_{r=p}^{q-1} k^r\right) d(x_1, x_0);$$

ce qui entraîne que :

$$d(x_q, x_p) \leqslant \frac{k^p}{1-k} d(x_1, x_0);$$

cette dernière expression tend vers 0 quand p tend vers $+\infty$. La suite (x_n) est donc de Cauchy et admet une limite x, puisque E est complet.

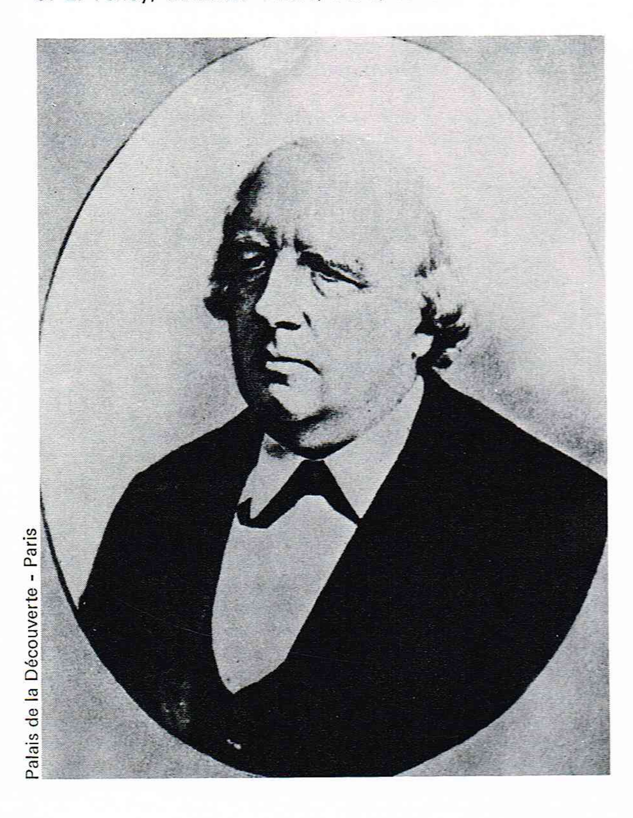
Remarquons que, si on fait tendre q vers l'infini dans l'inégalité précédente, on obtient :

$$d(x_q, x_p) \leqslant \frac{k^p}{1-k} d(x_1, x_0);$$

ce qui précise la rapidité de convergence de la suite (x_n) vers x.

BIBLIOGRAPHIE

BOURBAKI N., *Éléments de mathématiques : Topologie générale*, ch. I à IV, nouvelle édition, Hermann, Paris, 1971. - CHOQUET G., *Cours d'analyse*, t. II, *Topologie*, nouvelle édition, Masson, Paris, 1969. - DIEUDONNÉ J., *Éléments d'analyse*, t. I, *les Fondements de l'analyse moderne*, nouvelle édition, Gauthier-Villars, Paris, 1968. - SCHWARTZ L., *Analyse : Topologie générale et Analyse fonctionnelle*, Hermann, Paris, 1970. - WALLACE A. H., *Introduction à la topologie algébrique*, traduite par J.-L. Verley, Gauthier-Villars, Paris, 1973.



Palais de la Découverte - Paris

◄ **Karl Weierstrass** ***(1815-1897).***

ANALYSE

Le terme « analyse » est, pour tout utilisateur des mathématiques, pris au sens d'analyse infinitésimale; c'est-à-dire que d'emblée il pose les problèmes du continu, de l'infini et donc de limite. On dit souvent que l'analyse commence où l'algèbre s'arrête; cette image n'oublie certainement pas que le continu prend naissance avec la construction du corps des nombres réels $\mathbb{R}$ (Dedekind, Weierstrass et Heine-Canter) par sa propriété d'inextensibilité, de telle sorte que $\mathbb{R}$ se construit par des procédés de passage à la limite, tandis que $\mathbb{Q}$ ou $\mathbb{Z}$ (voir *les Nombres*) sont envisagés comme des constructions algébriques stables : anneaux, corps. L'utilisation de la topologie permet de préciser davantage ce point.

C'est donc bien par souci de distinction avec la synthèse algébrique que s'est développée l'analyse en mathématiques, à partir de la notion de continu telle que la mécanique, par exemple, avec l'idée de mouvement, de vitesse, d'accélération, la fournit.

Malgré l'extension du calcul infinitésimal depuis le XVII[e] siècle, ce n'est qu'à la fin du XIX[e] siècle avec les travaux de Georg Cantor (1845-1918) que les concepts pourront être délimités formellement tant pour le continu que pour l'infini. Le développement de la topologie (Riemann, Fréchet) donnera alors un cadre global à l'analyse mathématique classique.

La démarche du mathématicien tendant à plus de généralité en se posant la question de l'approximation linéaire des fonctions conduisit à construire le calcul différentiel sur les espaces de Banach, puis la géométrie différentielle — généralisant à plus de trois dimensions l'idée de Descartes — et aboutit à ce pseudo-paradoxe qu'est l'algébrisation de l'analyse, que l'on retrouve par exemple dans les théories modernes de la mesure (où l'on fait une harmonieuse synthèse de l'analyse et de l'algèbre par l'intermédiaire de la notion de dualité).

Au cœur de toute l'analyse mathématique se trouve la notion de correspondance « fonctionnelle » : pour l'évolution des grands courants d'idées mathématiques, c'est là une des plus fructueuses sources, et l'on peut encore y voir un lien étroit entre l'algèbre et la mathématique du continu. Cette interpénétration des domaines est encore affirmée par la théorie analytique des nombres premiers, ou bien d'autres travaux récents.

Analyse classique

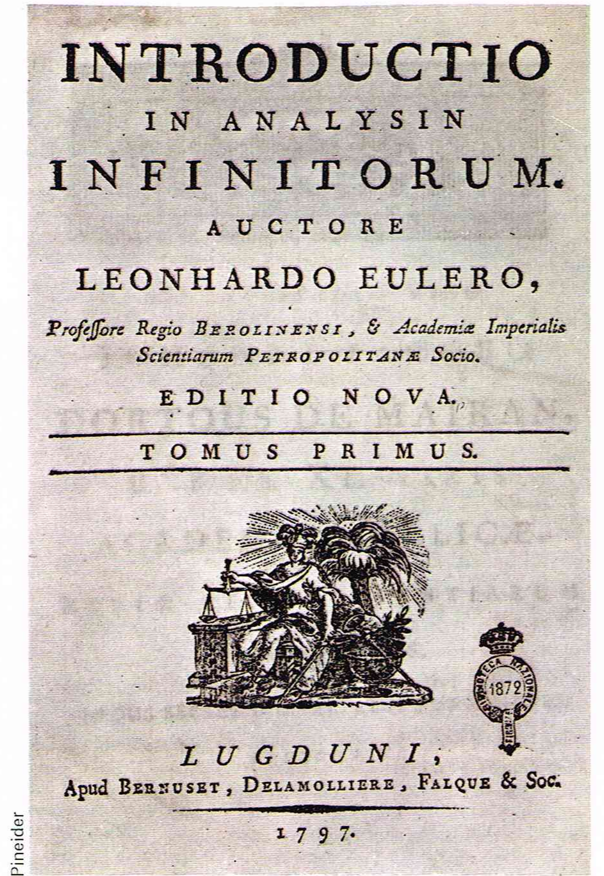
Variables et fonctions

Dans cette première partie, on va dresser un cadre général de l'analyse classique. On ne considère que des nombres réels, laissant le cas complexe pour un développement ultérieur. Les concepts seront dégagés dans leur sens le plus élémentaire afin de montrer la nécessité de généralisations.

Une *grandeur variable*, ou plus simplement une *variable*, est une grandeur susceptible de prendre plusieurs valeurs; une *grandeur constante*, ou encore une *constante*, garde en revanche une valeur donnée. Selon les conditions du problème posé, une même grandeur peut être soit variable, soit constante : par exemple, au cours de l'étude d'un phénomène physique, la température d'ébullition de l'eau est une constante dès que toutes les conditions sont spécifiées, mais elle devient une variable si l'on fait varier la pression ambiante.

C'est au mathématicien allemand Bernhardt Riemann (1826-1866) que l'on doit la définition claire d'une *fonction :* « Le nombre y est fonction du nombre x qui varie dans un champ donné G si, à tout nombre x situé dans D, correspond par une loi déterminée mais de nature et d'expression absolument quelconque un nombre y et un seul; x est appelé la variable indépendante. » Sa définition fait écho à celle de son compatriote Gustav Dirichlet (1805-1859) : « On dit que y est une fonction de x si, lorsqu'on se fixe une certaine valeur de x quelconque choisie dans un certain ensemble, on en déduit une valeur de y et une seule. »

Il s'agit là d'un tournant dans le développement de l'analyse mathématique prise en tant qu'étude des modes de variation simultanée de nombres liés entre eux. En effet, jusqu'au XVIII[e] siècle, on a pu se contenter d'une certaine notion de fonction, très ambiguë, car non séparée


INTRODUCTIO
IN ANALYSIN
INFINITORUM.
AUCTORE
LEONHARDO EULERO,
Professore Regio BEROLINENSI, & *Academiæ Imperialis Scientiarum* PETROPOLITANÆ *Socio.*
EDITIO NOVA.
TOMUS PRIMUS.
LUGDUNI,
Apud BERNUSET, DELAMOLLIERE, FALQUE & Soc.
1797.

Pineider

▶ *Frontispice de l'* « Introduction à l'analyse des infiniment petits » *de Leonhard Euler dans une édition, datée de 1797, de cette œuvre publiée pour la première fois en 1748.*

▶ *Page ci-contre en haut à gauche, figure 3 : représentation graphique correspondant à deux fonctions distinctes. En haut à droite, figure 6 :*

a, la fonction $y = \text{Arc cos}\sqrt{1 - t^2}$ *est définie et continue pour* $t \in [-1, +1]$*;*

b, la fonction $y = \begin{cases} -1 & \text{si } x < 0 \\ 0 & \text{si } x = 0 \\ +1 & \text{si } x > 0 \end{cases}$ *présente une discontinuité de première espèce en* $x = 0$*;*

c, la fonction $y = \sin\frac{1}{x}$ *présente une discontinuité de seconde espèce en* $x = 0$*;*

d, la fonction $y = \cos\frac{\pi}{x}$ *est définie et continue pour tout* $x \neq 0$*; on ne peut pas prolonger par continuité cette fonction en* $x = 0$*.*

d'un contexte analytique et incluant la continuité; mais après Leonhard Euler (1707-1783), dès les travaux de Jean-Baptiste-Joseph Fourier (1768-1830), et surtout ceux d'Augustin-Louis Cauchy (1789-1857), il devint nécessaire de donner une définition intrinsèque à la notion de correspondance fonctionnelle.

Par le tableau suivant, on a cherché à exprimer la dépendance entre la température T d'ébullition de l'eau et la pression atmosphérique *p* :

p (mm de mercure)	300	400	500	600	700
T (°C)	75,8	83,0	88,7	93,5	97,7

Sur la première ligne sont reportées diverses valeurs de la pression et, sur la seconde — en regard sur chaque colonne — les valeurs correspondantes de la température. On peut donc dire que la température T est fonction de la pression *p*; il est nécessaire, toutefois, d'ajouter quelques observations. Tout d'abord, on note que le tableau ne contient pas toutes les valeurs numériques que peut prendre la grandeur « pression ». En second lieu, on voit qu'à chaque valeur de la température, correspond une valeur bien déterminée de la pression ; on peut ainsi penser qu'il est possible d'échanger les rôles des deux lignes du tableau et de considérer la pression atmosphérique comme étant fonction de la température d'ébullition de l'eau.

Notons encore que, parmi les cas tabulés, la valeur de la température T = — 300 °C ne correspond à aucune valeur de la pression, non parce que, sur cette table, on ne peut lire la valeur — 300 °C, mais parce qu'une telle valeur de température n'est physiquement pas admissible (le zéro absolu étant proche de — 273 °C). Ceci éclaire le sens de l'expression « choisie dans un certain ensemble » utilisée dans la définition de G. Dirichlet pour le choix d'une valeur de la variable indépendante.

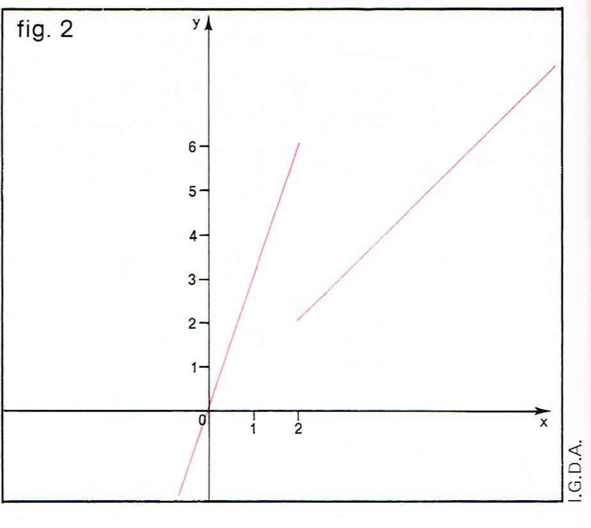
Définir une fonction signifie donc se donner la possibilité de connaître les valeurs prises par la variable qui dépend — selon une certaine correspondance — des valeurs prises par une autre variable, indépendante; cela peut se faire par un tableau (la spécification n'est alors que partielle), ou par un graphique en considérant les deux variables comme les coordonnées cartésiennes d'un point du plan (l'avantage est alors d'obtenir une idée globale du comportement de la fonction donnée), soit encore analytiquement, quand cela est possible, lorsque la fonction peut admettre une ou plusieurs « formules » (*fig.* 1 et 2).

Le *domaine de définition* d'une fonction définie par une ou plusieurs « formules » est l'ensemble des valeurs de la variable dépendante pour lesquelles ces « formules » ont un sens. Ainsi, la somme des angles internes d'un polygone est fonction du nombre de côtés; elle est donc définie seulement pour des valeurs entières et positives de la variable dépendante. La « formule »

$$y = \sqrt{x - 2} + \sqrt{7 - x}$$

n'a de sens (dans le champ réel) que pour $2 \leqslant x \leqslant 7$; le domaine de définition de cette fonction est donc l'intervalle fermé [2, 7].

On distingue souvent les *fonctions algébriques*, qui se calculent par des opérations rationnelles et des extractions de racines, et les *fonctions transcendantes*, qui sont les fonctions analytiques non algébriques; parmi ces dernières, on peut citer les fonctions exponentielle, logarithme, trigonométriques directes, et bien d'autres encore. Cette classification n'est, bien sûr, pas exhaustive, et de nombreuses fonctions ne rentrent pas dans le cadre des fonctions analytiques, ainsi la fonction indicatrice des nombres irrationnels :



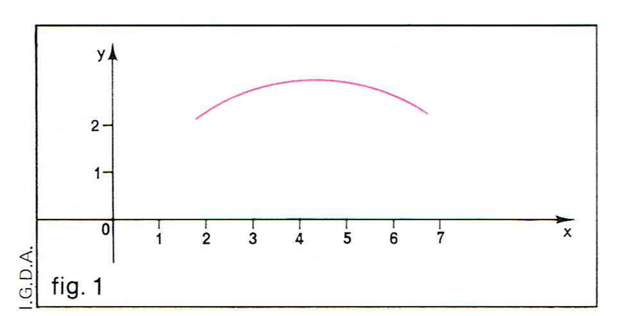
I.G.D.A.

▶ *Figure 2 : graphique de la fonction définie par* $y = 3x$ *pour* $x < 2$*,* $y = x$ *pour* $x > 2$*.*

$$\begin{cases} y = 1 & \text{si } x \notin \mathbb{Q} \\ y = 0 & \text{si } x \in \mathbb{Q} \end{cases}$$

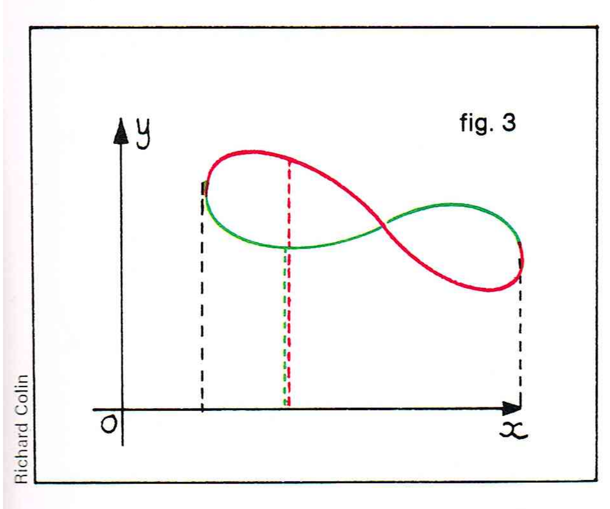
Par la représentation graphique des fonctions, on peut faire ressortir une propriété de définition : le graphe d'une fonction n'est coupé par une parallèle quelconque à l'axe des ordonnées O*y* qu'en, *au plus*, un point. La figure 3 montre ainsi que la courbe tracée correspond à autant de fonctions distinctes que d'arcs possédant la propriété ci-dessus, c'est-à-dire deux.

Ce sont les propriétés des nombres réels que l'on retrouve dans les caractéristiques de variation des fonctions : structure d'ordre, bornes d'intervalles, que l'on



I.G.D.A.

▶ *Figure 1 : graphique de la fonction* $y = \sqrt{x - 2} + \sqrt{7 - x}$*, pour* $2 \leqslant x \leqslant 7$*.*



Richard Colin

traduit en termes de croissance et d'extrema. Dans ce qui suit, on notera $f(x)$ le nombre associé à la variable x par la correspondance ou fonction $x \mapsto f(x)$.

On dit qu'une fonction est *croissante* sur un domaine D, inclus dans le domaine de définition, si :

$$\forall x_1 \in D, x_2 \in D, x_1 \leqslant x_2 \Rightarrow f(x_1) \leqslant f(x_2).$$

Elle est dite *décroissante* sur D si :

$$\forall x_1 \in D, x_2 \in D, x_1 \leqslant x_2 \Rightarrow f(x_1) \geqslant f(x_2).$$

Elle est dite *monotone* sur D si elle est croissante ou décroissante sur D (*fig.* 4).

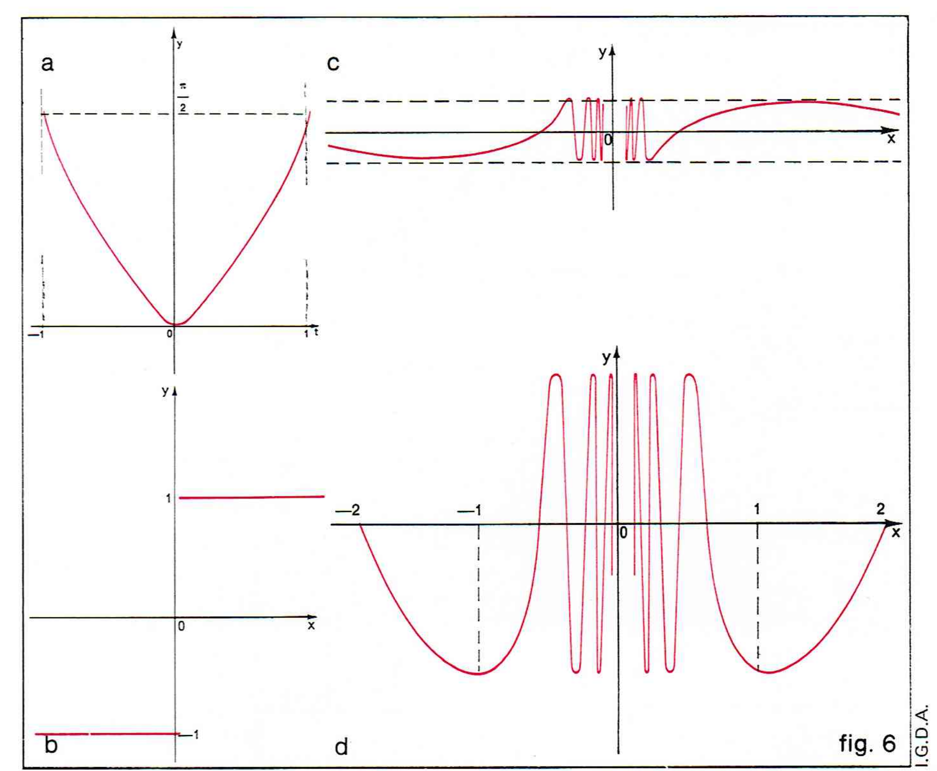
On dit qu'une fonction est *paire* si $f(x) = f(-x)$ pour tout $x \in D$, et qu'elle est *impaire* si $f(x) = -f(-x)$ pour tout $x \in D$ (*fig.* 5). La représentation graphique admet dans le premier cas O*y* pour axe de symétrie, dans le second cas l'origine O pour centre de symétrie.

La notion de continuité

La topologie induite sur un ensemble par une métrique a permis de définir avec une grande netteté l'idée de *limite* et celle de *continuité* d'une fonction.

On dit que la fonction définie sur l'ensemble $E \subseteq \mathbb{R}$ admet la *limite* L lorsque x tend vers a si, à tout nombre $\varepsilon > 0$ aussi petit qu'on le désire, on peut associer un nombre η — dépendant de ε — tel que, lorsque $|x - a| < \eta$ et $x \in E$, alors $|f(x) - L| < \varepsilon$, ce que l'on écrit $\lim f(x) = L$. En particulier, la fonction est *continue* au point $a \in E$ si L existe et est égal à $f(a)$. On appelle *points de discontinuité* de la fonction les points du domaine de définition en lesquels elle n'est pas continue. Ceux-ci sont de trois types :

tout d'abord les *points à discontinuité éliminable*, par exemple l'origine pour la fonction définie par :



I.G.D.A.

$$y = \sin x \quad \text{si } x \neq 0 \qquad \text{et} \qquad y = 2 \quad \text{si } x = 0$$

puisqu'il suffit de changer la valeur en $x = 0$ en posant $y = 0$ pour obtenir la continuité ;

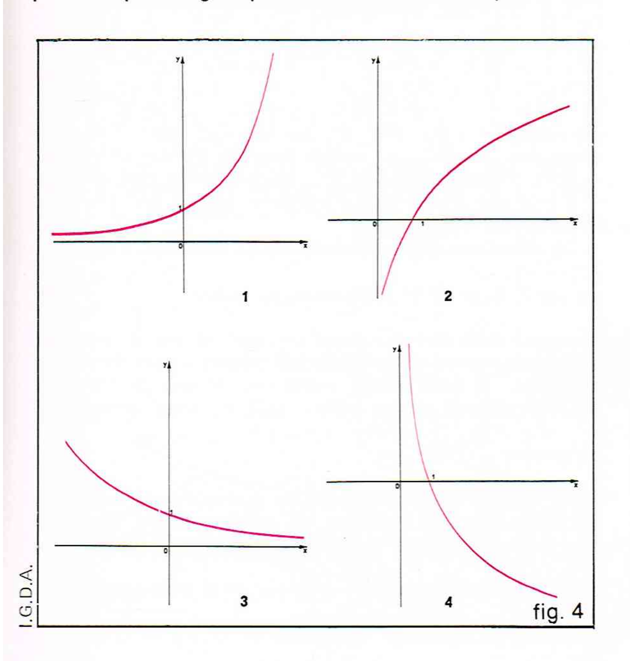
ensuite les *points de discontinuité de première espèce* pour lesquels existent une limite à droite L, une limite à gauche L', et tels que la valeur en ces points soit égale à $\frac{1}{2}(L + L')$; par exemple, la fonction définie par

$$y = \frac{\sqrt{x^2}}{x} \quad \text{si } x \neq 0 \qquad \text{et} \qquad y = 0 \quad \text{si } x = 0 ;$$

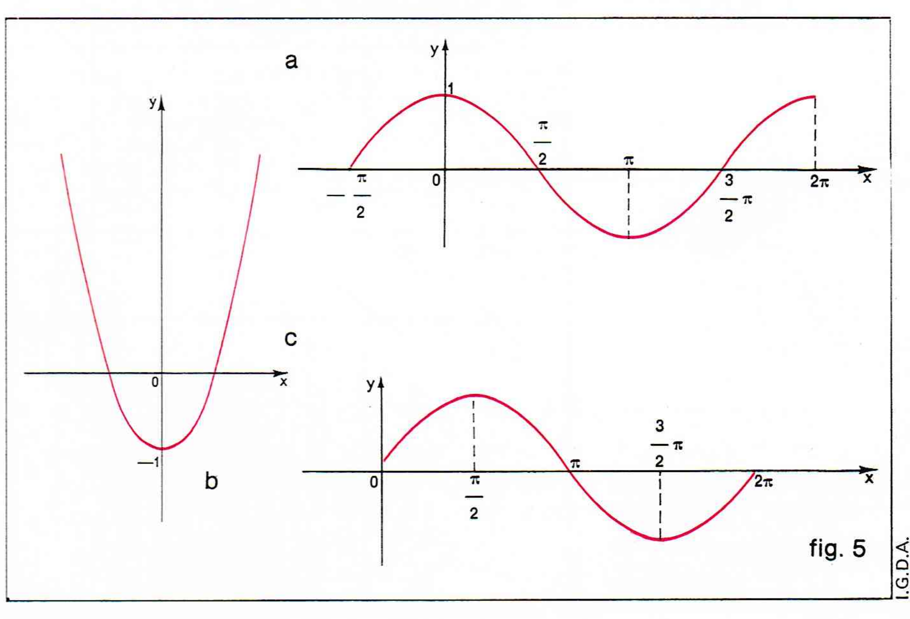
enfin, les *points de discontinuité de seconde espèce*, tels que, soit la limite à gauche, soit la limite à droite, soit les deux, n'existent pas ou bien soient infinies (*fig.* 6).

On peut rappeler qu'une condition nécessaire et suffisante pour que la fonction strictement monotone $x \mapsto f(x)$ définie sur l'intervalle $[a, b]$ soit continue sur cet intervalle, est que cette fonction prenne toutes les valeurs comprises entre $f(a)$ et $f(b)$. Ceci permet entre autres de définir les fonctions réciproques des fonctions

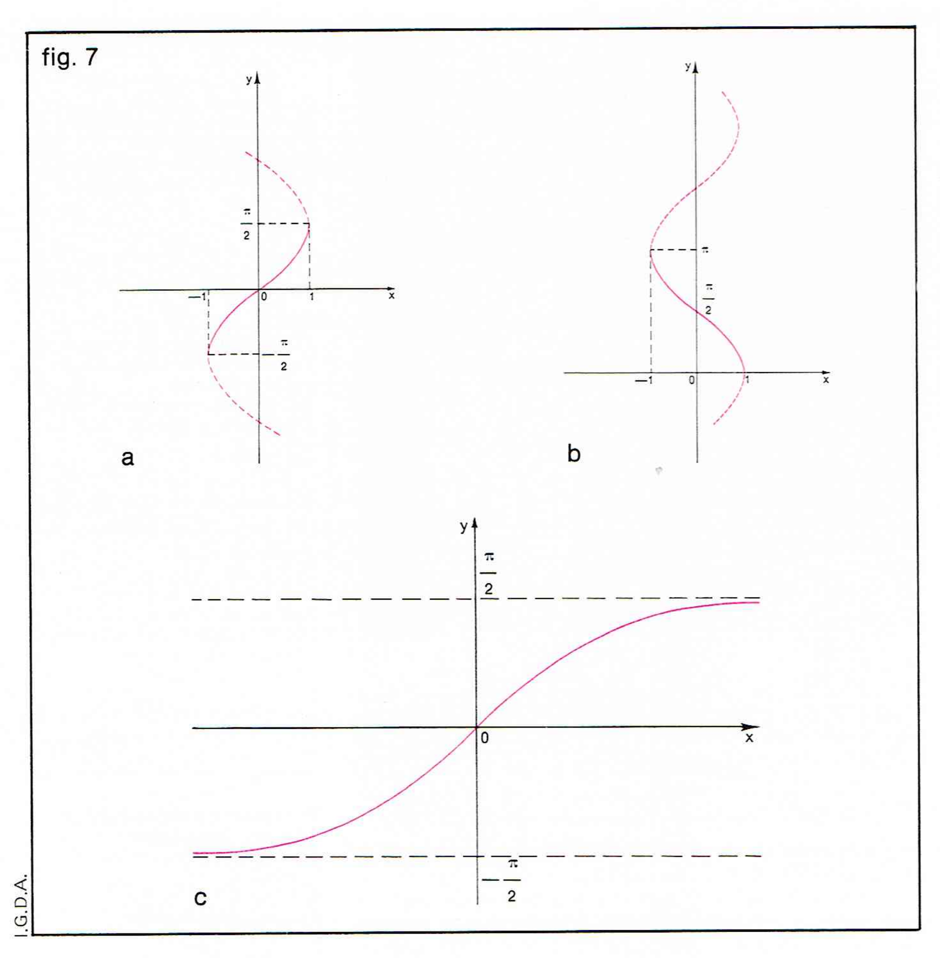
▼ A gauche, figure 4 : en haut, graphiques de la fonction** $y = a^x$ **: 1, pour** $a > 1$ **; 2, pour** $0 < a < 1$ **; en bas, graphiques de la fonction** $y = \log_a(x)$ **; 3, pour** $a > 1$ **; 4, pour** $0 < a < 1$**. Les graphiques 1 et 2 sont ceux des fonctions croissantes, alors que les graphiques 3 et 4 sont ceux de fonctions décroissantes. A droite, figure 5 : a, un exemple de fonction paire,** $y = \cos x$ **; b, un autre exemple de fonction paire,** $y = 3x^2 - 1$ **; c, un exemple de fonction impaire,** $y = \sin x$**.

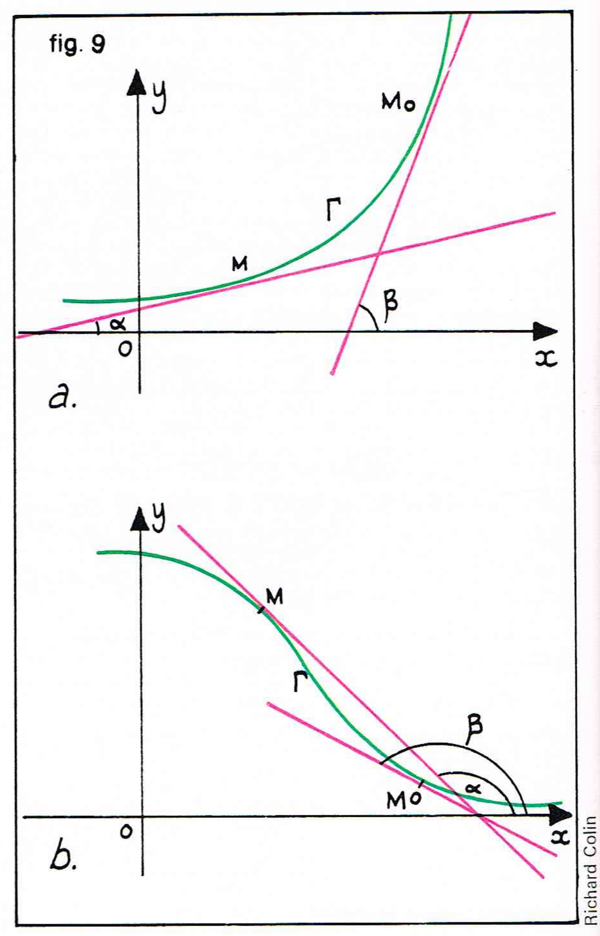


I.G.D.A.



I.G.D.A.





▲ *A gauche, figure 7 :*
***a**, graphique de la fonction*
y = Arc sin x;
***b**, celui de la fonction*
y = Arc cos x;
***c**, celui de la fonction*
y = Arc tg x.
A droite, figure 9 représentant une fonction croissante et une fonction décroissante :
***a**, la croissance est plus forte en M_0 qu'en M car tg $\beta >$ tg α;*
***b**, la décroissance est plus forte en M qu'en M_0 car tg $\alpha >$ tg β.*

usuelles continues et strictement monotones sur un intervalle (*fig.* 7).

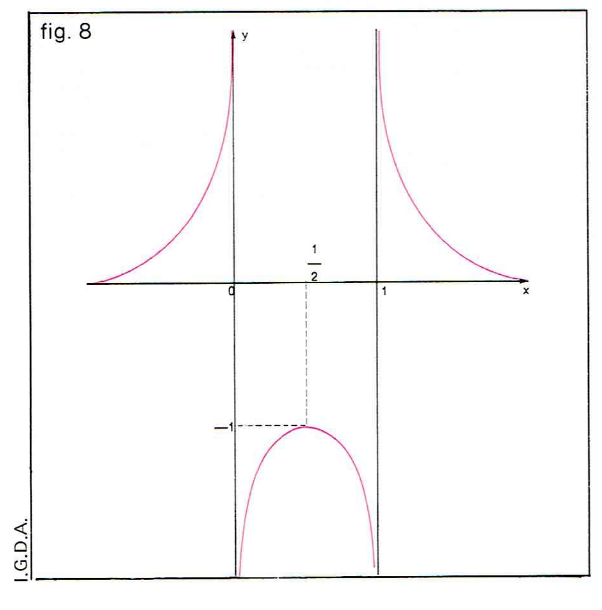
Citons parmi les principales propriétés des fonctions continues sur un intervalle I fermé borné de $\mathbb{R}$:

a — toute fonction réelle continue dans I est bornée dans I ;

b — toute fonction réelle continue dans I possède un minimum et un maximum dans I, qui sont effectivement atteints par la fonction ;

c — une fonction continue dans I ne peut y changer de signe que si elle s'y annule ;

d — toute fonction réelle, continue dans I, y est *uniformément continue*, c'est-à-dire que, pour tout $\varepsilon > 0$, il existe $\eta > 0$, ne dépendant *que de* ε, tel que, si x' et x'' sont des points de I tels que $|x' - x''| \leqslant \eta$, alors $|f(x') - f(x'')| \leqslant \varepsilon$.



▶ *Figure 8 :*
graphique de la fonction
$y = \dfrac{1}{4x(x-1)}$ *continue sur]0, 1[, mais non bornée.*

Les propriétés **a**, **b** et **d** ne sont plus vraies dès que I n'est pas fermé ou n'est pas borné (*fig.* 8).

Dérivées et primitives

L'idée même de la dérivée d'une fonction est très couramment répandue, puisque c'est en fait celle de vitesse d'un mouvement, ou bien celle de pente d'une route, ou même encore celle de moyenne. C'est une considération géométrique qui a permis de développer ce point, et par là même, de construire l'analyse mathématique ; on doit cette étape fondamentale au génial mathématicien allemand Gottfried Wilhelm Leibniz (1646-1716).

Sur une courbe Γ parcourue par un mobile M (ou décrite par un point M) dans le sens des abscisses croissantes, on montera ou l'on descendra, aux alentours d'un point M_0, si la fonction représentée par Γ est croissante ou décroissante, et l'on constatera aisément que ce phénomène de montée ou de descente est d'autant plus accentué que la pente de la tangente Γ en M_0 est élevée en valeur absolue (*fig.* 9). De là vient donc l'idée de considérer non plus les cordes joignant deux points de Γ, mais la tangente au point fixe, dont la pente sera la limite — lorsqu'elle existe — de la pente de la corde M_0M lorsque M tend vers M_0 (*fig.* 10).

On appellera donc *dérivée* de la fonction $x \mapsto f(x)$ au point x_0 la limite éventuelle du rapport $\dfrac{f(x) - f(x_0)}{x - x_0}$ lorsque x tend vers x_0, ce qui correspond bien à la notion usuelle de vitesse instantanée par rapport à celle de vitesse moyenne. Si cette limite existe, on dit que la fonction est *dérivable* en x_0 ; on note $f'(x_0)$ la valeur de celle-ci, ou encore $\dfrac{df}{dx}(x_0)$.

On montre que toute fonction dérivable est continue. La réciproque est inexacte : le mathématicien allemand Karl Weierstrass (1815-1897) a par exemple montré que la fonction définie par : $\sum_{n=0}^{\infty} a^n \cos \pi\, b^n x$ était continue et

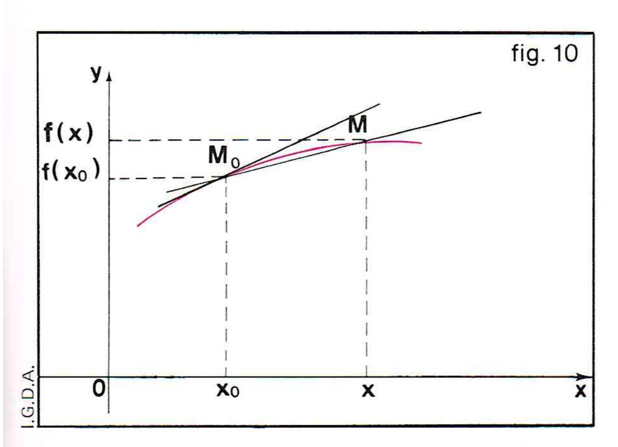


Tableau 11 - Dérivées usuelles

Fonction	Dérivée	Fonction	Dérivée
c^{te}	0	f + g	f' + g'
x^n	$n\,x^{n-1}$		
Log x	$\frac{1}{x}$	af	af'
sin ax	a cos ax	fg	f'g + fg'
cos ax	− a sin ax	$\frac{f}{g}$	$\frac{f'g - fg'}{g^2}$
tg ax	$\frac{a}{\cos^2 ax}$	f ∘ g	f'g'
e^{ax}	ae^{ax}	f^n	$n\,f^{n-1}f'$
Arc sin x	$\frac{1}{\sqrt{1-x^2}}$	Log| f |	$\frac{f'}{f}$
Arc cos x	$\frac{-1}{\sqrt{1-x^2}}$	e^f	$f'e^f$
Arc tg x	$\frac{1}{1+x^2}$	sin f	f' cos f

non dérivable, si $0 < a < 1$ et $b \in \mathbb{N}$ tel que $ab > \frac{3\pi}{2} + 1$.

Lorsqu'une fonction $x \mapsto f(x)$ admet une dérivée (ou est dérivable) en tout point d'un domaine D, on peut donc associer à tout point $x \in$ D le nombre $f'(x)$ et définir ainsi une nouvelle fonction dite fonction dérivée $x \mapsto f'(x)$.

L'opération qui consiste, partant d'une fonction donnée $f(x)$, à en déduire la fonction dérivée $f'(x)$, s'appelle donc la *dérivation*, et le tableau 11 résume les résultats les plus notables. Cette opération peut être répétée à nouveau, en partant de la fonction $f'(x)$; on obtient ainsi la fonction *dérivée seconde* $x \mapsto f''(x)$, puis la fonction *dérivée troisième*, et ainsi de suite ; la *dérivée n-ième* $f^{(n)}(x)$ étant donc la dérivée de la dérivée d'ordre $(n-1)$, $f^{(n-1)}(x)$.

La dérivation des fonctions fait partie de la classe des outils fondamentaux de l'analyse mathématique, car elle va permettre de définir une méthode générale pour étudier localement (au voisinage d'un point) le comportement d'une fonction en la remplaçant — lorsque cela est nécessaire — par une fonction plus simple.

On a vu dans les considérations géométriques qui introduisaient l'idée de dérivée que, plus une fonction croissait ou décroissait, plus la pente de la tangente au graphe était élevée en valeur absolue. D'autre part, il est clair que cette pente est négative si la fonction décroît, positive si elle croît. On peut donc se poser la question de savoir ce qui se passe pour un extremum — minimum ou maximum — c'est-à-dire en un point où l'on passe d'un sens de variation à l'autre (il s'agit d'extrema relatifs) ; on retrouve alors que la dérivée est nulle en un tel point et change de signe.

Le **théorème de Rolle** ne fait que préciser ceci en disant qu'une fonction continue sur un intervalle fermé $[a, b]$, et dérivable sur $]a, b[$, telle que $f(a) = f(b)$, admet nécessairement un extremum (au moins) dans $[a, b]$ (*fig.* 12) : en effet, partant du point $(a, f(a))$ en croissant, pour revenir à la même hauteur il faudra « descendre » à un moment donné ; de même, si l'on décroît à partir de $(a, f(a))$, il faudra « remonter ». Si la corde qui joint les extrémités du graphe représentant $x \mapsto f(x)$ sur $[a, b]$ est horizontale, il en sera donc de même de la tangente au graphe en un point (au moins).

C'est cette présentation qui permet de voir que le **théorème des accroissements finis** est la généralisation du théorème de Rolle : si $f(x)$ est continu sur $[a, b]$, et dérivable sur $]a, b[$, alors il existe un point $c \in]a, b[$ tel que $f(b) - f(a) = (b - a)\, f'(c)$. En d'autres termes, il existe un point c où la tangente est parallèle à la corde joignant $(a, f(a))$ et $(b, f(b))$ (*fig.* 13), puisqu'en ce point $f'(c) = \frac{f(b) - f(a)}{b - a}$. Il suffit d'appliquer le théorème de Rolle à la fonction

$$F(x) = f(x) - f(a) - \frac{f(b) - f(a)}{b - a}(x - a)$$

qui est telle que $F(b) = F(a) = 0$, et l'on a donc un point $c \in]a, b[$ où $F'(c) = 0$, or

$$F'(c) = f'(c) - \frac{f(b) - f(a)}{b - a}$$

Lorsqu'on écrit la formule des accroissements finis sous la forme $f(b) = f(a) + (b - a)\, f'(c)$, on voit apparaître l'idée qui est exprimée par la *formule de Taylor* (Brook Taylor, 1685-1731, mathématicien anglais), qui consiste à approcher les valeurs d'une fonction autour d'un point a par un polynôme selon les puissances de la différence $(x - a)$; en effet, dans la formule des accroissements finis, en supposant la dérivée continue et b très voisin de a, alors $f'(b)$ est très voisin de $f'(a)$:

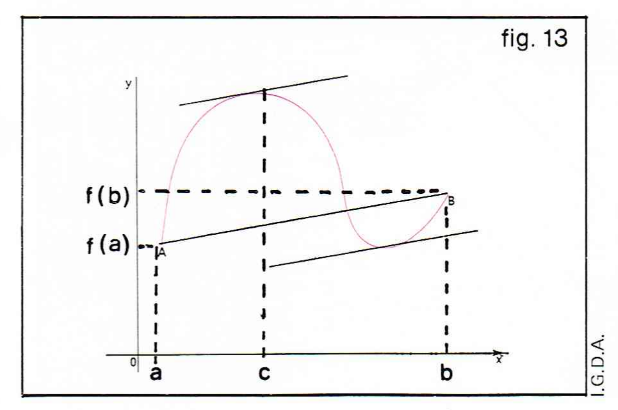
$$f'(b) = f'(a) + (b - a)\,\varepsilon,$$

où ε tend vers 0 lorsque $b \to a$. On obtient alors

$$f(b) = (b - a)\, f'(a) + (b - a)^2\,\varepsilon.$$

Il suffit alors de poursuivre ce procédé en l'appliquant aux dérivées successives tant qu'elles vérifient les hypothèses de départ. On obtient alors que, pour toute fonction $x \mapsto f(x)$ définie et continue sur $[a, b]$, dont les n premières dérivées existent et sont continues sur $]a, b[$, et $(n + 1)$ fois dérivable au point a, on peut écrire :

▲ *A gauche, figure 10 :* ***représentation graphique du concept de dérivée.*** *A droite, tableau 11 :* ***dérivées des fonctions les plus usuelles et règles d'opérations sur les dérivées.***



◀ *A gauche, figure 12 :* ***interprétation graphique du théorème de Rolle.*** *A droite, figure 13 :* ***interprétation graphique du théorème des accroissements finis.***

Tableau 14 - Développement en série des fonctions analytiques usuelles

Fonction	Développement en série
e^x	$1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!} + \cdots$
$\sin x$	$x - \frac{x^3}{3!} + \frac{x^5}{5!} + \cdots + (-1)^n \frac{x^{2n+1}}{(2n+1)!} + \cdots$
$\cos x$	$1 - \frac{x^2}{2!} + \frac{x^4}{4!} + \cdots + (-1)^n \frac{x^{2n}}{(2n)!} + \cdots$
$(1+x)^m$	$1 + mx + \frac{m(m-1)}{2!}x^2 + \cdots + \frac{m(m-1)\ldots(m-p+1)}{p!}x^p + \cdots$ si $\lvert x \rvert \leqslant 1$
$\mathrm{Log}\,(1+x)$	$x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots + (-1)^{n+1}\frac{x^n}{n} + \cdots$ si $x \in]-1, +1]$
$\mathrm{Log}\,(1-x)$	$-\left(x + \frac{x^2}{2} + \frac{x^3}{3} + \cdots + \frac{x^n}{n} + \cdots\right)$ si $x \in [-1, +1[$
$\mathrm{Arc\,sin}\,x$	$x + \frac{x^3}{2.3} + \frac{1.3\,x^5}{2.4.5} + \frac{1.3.5\,x^7}{2.4.6.7} + \cdots + \frac{1.3.\ldots(2n-1)x^{2n+1}}{2.4.6.\ldots(2n)(2n+1)} + \cdots$ si $\lvert x \rvert < 1$
$\mathrm{Arc\,tg}\,x$	$x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \cdots + (-1)^n \frac{x^{2n+1}}{2n+1} + \cdots$ si $\lvert x \rvert < 1$

Tableau 15 - Recherche de primitives

Fonction	Primitive	Fonction	Primitive	Fonction	Primitive
$(ax+b)^p$	$\frac{1}{a(p+1)}(ax+b)^{p+1}$ si $p \neq -1$	$\sqrt{a^2+x^2}$	$\frac{1}{2}\left[x\sqrt{a^2+x^2} + a^2 \mathrm{Log}\,(x+\sqrt{a^2+x^2})\right]$	$\sin^n ax \cos ax$	$\frac{1}{a(n+1)}\sin^{n+1} ax$ si $n \neq -1$
$\frac{1}{ax+b}$	$\frac{1}{a}\mathrm{Log}\,(ax+b)$	$x\sqrt{a^2+x^2}$	$\frac{1}{3}\sqrt{(a^2+x^2)^3}$	$\cos^n ax \sin ax$	$\frac{-1}{a(n+1)}\cos^{n+1} ax$ si $n \neq -1$
$\frac{ax+b}{cx+d}$	$\frac{ax}{c} + \frac{bc-ad}{c^2}\mathrm{Log}\,(cx+d)$	$\sqrt{a^2-x^2}$	$\frac{1}{2}\left[x\sqrt{a^2-x^2} + a^2 \mathrm{Arc\,sin}\,\frac{x}{a}\right]$	$\mathrm{tg}\,ax$	$-\frac{1}{a}\mathrm{Log}\cos ax$
$\frac{1}{ax^2+bx+c}$	$\frac{2}{\sqrt{4ac-b^2}}\mathrm{Arc\,tg}\,\frac{2ax+b}{\sqrt{4ac-b^2}}$ si $4ac-b^2>0$	$\frac{1}{\sqrt{a^2-x^2}}$	$\mathrm{Arc\,sin}\,\frac{x}{a}$	e^{ax}	$\frac{1}{a}e^{ax}$
	$\frac{-2}{\sqrt{b^2-4ac}}\mathrm{Arg\,th}\,\frac{2ax+b}{\sqrt{b^2-4ac}}$ si $b^2-4ac>0$	$\frac{1}{(ax^2+bx+c)^{3/2}}$	$\frac{2(2ax+b)}{(4ac-b^2)\sqrt{ax^2+bx+c}}$	xe^{ax}	$\frac{e^{ax}}{a^2}(ax-1)$
$\frac{1}{a^2+x^2}$	$\frac{1}{a}\mathrm{Arc\,tg}\,\frac{x}{a}$	$\sqrt[n]{ax+b}$	$\frac{n(ax+b)}{(n+1)a}\sqrt[n]{ax+b}$	$\mathrm{Log}\,x$	$x\,\mathrm{Log}\,x - x$
$\frac{1}{a^2-x^2}$	$\frac{1}{2a}\mathrm{Arg\,th}\,\frac{x}{a}$	$\sin ax$	$-\frac{1}{a}\cos ax$	$(\mathrm{Log}\,x)^2$	$x(\mathrm{Log}\,x)^2 - 2x\,\mathrm{Log}\,x + 2x$
$\frac{1}{\sqrt{ax+b}}$	$\frac{2}{a}\sqrt{ax+b}$	$\sin^2 ax$	$\frac{1}{2}x - \frac{1}{4a}\sin 2ax$	$\frac{(\mathrm{Log}\,x)^n}{x}$	$\frac{(\mathrm{Log}\,x)^{n+1}}{n+1}$ si $n \neq -1$
$\frac{x}{\sqrt{ax+b}}$	$\frac{2(ax-2b)}{3a^2}\sqrt{ax+b}$	$x \sin ax$	$\frac{\sin ax}{a^2} - \frac{x\cos ax}{a}$	$x^m \mathrm{Log}\,x$	$x^{m+1}\left[\frac{\mathrm{Log}\,x}{m+1} - \frac{1}{(m+1)^2}\right]$ si $m \neq -1$
$\sqrt{(ax+b)^n}$	$\frac{2\sqrt{(ax+b)^{n+2}}}{a(n+2)}$	$\cos ax$	$\frac{1}{a}\sin ax$	$\sin(\mathrm{Log}\,x)$	$\frac{x}{2}(\sin \mathrm{Log}\,x - \cos \mathrm{Log}\,x)$
$\frac{x}{a^4+x^4}$	$\frac{1}{2a^2}\mathrm{Arc\,tg}\,\frac{x^2}{a^2}$	$\sin ax \cos ax$	$\frac{1}{2a}\sin^2 ax$	$e^{ax}\sin bx$	$\frac{e^{ax}}{a^2+b^2}(a \sin bx - b\cos bx)$
$\frac{x}{a^4-x^4}$	$\frac{1}{4a^3}\mathrm{Log}\,\frac{a^2+x^2}{a^2-x^2}$	$\frac{1}{\sin ax \cos ax}$	$\frac{1}{a}\mathrm{Log}\,\mathrm{tg}\,ax$	$\mathrm{Arc\,sin}\,\frac{x}{a}$	$x\,\mathrm{Arc\,sin}\,\frac{x}{a} + \sqrt{a^2-x^2}$

$$f(b) = f(a) + \frac{(b-a)}{1!}f'(a) + \frac{(b-a)^2}{2!}f''(a) + \ldots + \frac{(b-a)^n}{n!}f^{(n)}(a) + \frac{(b-a)^{n+1}}{(n+1)!}[f^{(n+1)}(a) + \varepsilon]$$

où $\varepsilon \to 0$ lorsque $b \to a$.

La même formule où l'on pose $x = b$ et $a = 0$ est appelée *formule de MacLaurin* (Colin MacLaurin, 1698-1746, mathématicien écossais) et pose le problème suivant : si la fonction f est indéfiniment dérivable, peut-on poursuivre le développement du second membre à un nombre infini de termes (afin d'obtenir une excellente précision d'approximation) ? En effet, la série alors obtenue est-elle convergente, et sa somme est-elle bien égale au premier membre ? La réponse est négative ; seule une certaine catégorie de fonctions, dites alors *analytiques*, possèdent cette propriété (voir *Séries* et *Fonctions de variable complexe*). Le tableau 14 donne les exemples les plus courants de ces fonctions et leur développement en série entière.

Le problème inverse de la dérivation des fonctions est celui de la recherche des primitives. On dit que la fonction F est une *primitive* de la fonction f sur l'intervalle $[a, b]$ si, en tout point de $[a, b]$, F est dérivable, et sa dérivée égale à f. On parle d'une des primitives de la fonction f, car, si F est une primitive, alors $F + k$, où k est constant, vérifie les mêmes propriétés. Il existe donc une infinité de primitives d'une fonction dès lors qu'il en existe une. On appelle alors *intégrale indéfinie* de la fonction $x \mapsto f(x)$ sur $[a, b]$ l'ensemble des primitives de la fonction sur $[a, b]$, et on note $\int f(x)\,dx$; le signe $\int$ est dit signe d'intégration (ou signe somme).

L'intégration est toutefois à séparer sur deux points de la dérivation :

— une fonction qui admet des primitives (on montre qu'il suffit qu'elle soit continue) n'est pas toujours dérivable ;

— le résultat de l'intégration n'est pas déterminé de manière unique.

Sur le tableau 15, on a figuré les résultats les plus courants de la recherche de primitives (que l'on pourra rapprocher du tableau 11).

Les liens entre cette notion, celle d'aire et celle d'intégrale définie sont développés plus loin, sous le titre *Intégration des fonctions*. Rappelons simplement que l'on appelle figure plane simple une réunion finie de triangles d'un même plan et que ces figures possèdent les propriétés de monotonie, additivité et invariance (voir *Géométrie*).

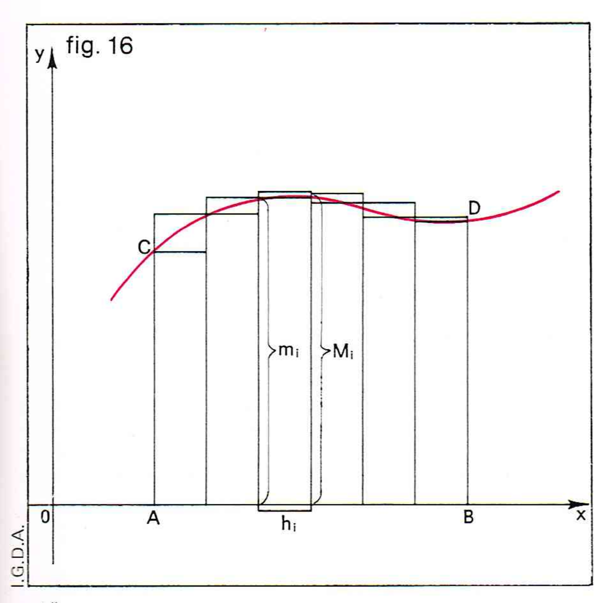
Pour une figure plane F quelconque, on considère l'ensemble des figures simples contenues dans F, soit $\mathcal{E}_1$, et l'ensemble des figures simples contenant F, soit $\mathcal{E}_2$. S'il existe, pour tout nombre ε aussi petit qu'on le désire, un élément de $\mathcal{E}_1$ et un élément de $\mathcal{E}_2$ tels que la différence de leurs surfaces soit inférieure à ε, on dit que la figure F est *mesurable au sens de Jordan ;* en posant alors $\mathcal{A}_1$, ensemble des nombres, surfaces des éléments de $\mathcal{E}_1$, et $\mathcal{A}_2$ ensemble des nombres, surfaces des éléments de $\mathcal{E}_2$, la surface ou aire de F sera le nombre séparant $\mathcal{A}_1$ et $\mathcal{A}_2$. L'aire d'une surface (*fig.* 16) définie par un arc de courbe est alors définie par la limite — lorsqu'elle existe — des sommes de Darboux :

$$\sum_{i=1}^{n-1}(x_{i+1} - x_i)\,f(x_i) \quad \text{et} \quad \sum_{i=1}^{n-1}(x_{i+1} - x_i)\,f(x_{i+1})$$

lorsque $n \to \infty$; ces sommes représentent la surface des rectangles inférieurs et supérieurs à la courbe. Lorsque la surface est mesurable, alors la limite existe et se note

$\int_a^b f(x)\,dx$ (intégrale définie de $f(x)$ sur l'intervalle $[a, b]$).

On peut alors montrer qu'en maintenant a fixe, et en faisant varier b — qu'on pose égal à une variable t — l'aire $\int_a^t f(x)\,dx = S(t)$, si elle existe en tout point t d'un intervalle I, est alors une fonction dérivable de t, et l'on a $S'(t) = f(t)$. En d'autres termes, toute fonction continue $f(x)$ définie sur un intervalle $[a, b]$ admet une primitive sur cet intervalle, qui n'est autre que



$\int_c^x f(t)\,dt$, où c est un point quelconque de $[a, b]$.

Il est alors aisé d'en déduire que, si F désigne une primitive de f sur l'intervalle $[a, b]$, on a

$$\int_a^b f(x)\,dx = F(b) - F(a),$$

formule permettant la réalisation des calculs.

La restriction sévère que l'on a mise à ce développement en ne formulant de résultats que pour les fonctions continues est levée dans le paragraphe *Intégration des fonctions.*

Fonctions de plusieurs variables

L'extension la plus naturelle des notions de calcul différentiel et intégral introduites pour les fonctions réelles d'une variable réelle est obtenue par l'étude des fonctions réelles de plusieurs variables réelles, c'est-à-dire des applications $f: \mathbb{R}^n \to \mathbb{R}$, que l'on représente par l'écriture $y = f(x_1, x_2, ..., x_n)$, où chaque x_i est une variable réelle; nous noterons parfois X le point $(x_1, x_2, ..., x_n) \in \mathbb{R}^n$, et donc $y = f(X)$. Il n'est pas nécessaire de développer plus particulièrement pour les fonctions de plusieurs variables les notions déjà dégagées de fonction et de domaine de définitions. Dans ce qui suit, on suppose que les variables $x_1, x_2, ..., x_n$ sont indépendantes, c'est-à-dire qu'il n'existe pas de correspondance présupposée entre elles.

Les topologies que l'on peut déterminer sur $\mathbb{R}^n$ permettent de définir les notions de limite et de continuité; par exemple, soit $y = f(X)$ une fonction définie sur un ensemble $E \subseteq \mathbb{R}^n$ et soit $X^0 = (x_1^0, x^0, ..., x_n^0)$ un point non isolé de E, on dit que $f(X)$ est *continue en* X^0 si, à tout nombre $\varepsilon > 0$ arbitraire, on peut associer un nombre $\eta > 0$ — en général η dépend de ε — tel que $|f(X) - f(X^0)| < \varepsilon$ dès que $X \in E$ est à distance de X^0 inférieure à η.

On définit aussi la continuité partielle par rapport à l'une des n variables réelles x_i en supposant fixes — donc constantes — les $(n-1)$ autres variables; y est à ce moment fonction d'une seule variable réelle. La continuité par rapport à l'ensemble des n variables réelles est une notion très forte qui inclut chaque continuité partielle; tandis qu'une fonction continue par rapport à chaque variable n'est pas nécessairement continue « globalement », cette dernière notion entraînant, en fait, la continuité selon n'importe quelle direction de $\mathbb{R}^n$. Par exemple, la fonction définie sur $\mathbb{R}^2$ par $y = \frac{x_1 x_2}{x_1^2 + x_1^2}$ si $(x_1, x_2) \neq (0,0)$ et $y = 0$ si $(x_1, x_2) = (0, 0)$ est continue par rapport à x_1 et par rapport à x_2, mais n'est pas continue le long de la bissectrice $x_1 = x_2$ puisque alors $y = \frac{1}{2}$ si $(x_1, x_2) \neq (0, 0)$, tandis que $y = 0$ si $(x_1, x_2) = (0, 0)$; cette fonction n'est donc pas continue à l'origine.

L'influence de chaque variable sur la variation de la fonction est mesurée par la *dérivée partielle de* f *par rapport à cette variable :* on appelle dérivée partielle de f par rapport à la variable x_k^0, au point X^0, la dérivée de la fonction de x_k, en x_k^0, obtenue en fixant les autres variables à leur valeur au point X^0 ; on note le résultat $\left(\frac{\partial f}{\partial x_k}\right)_{X^0}$, ce symbole n'étant évidemment pas une fraction.

Plus généralement, on appelle dérivée le long de la direction V (V étant un vecteur unitaire issu du point X^0) la limite du rapport $\frac{f(X^0 + hV) - f(X^0)}{h}$ lorsque $h \in \mathbb{R}$ tend vers zéro de telle sorte que $X^0 + hV$ reste dans le domaine de définition de f. Il faut noter que, contrairement à ce qui se passe dans le cas des fonctions d'une variable réelle, l'existence des dérivées partielles n'entraîne pas la continuité; l'existence de la dérivée partielle par rapport à une variable entraîne la continuité partielle par rapport à celle-ci. Toutefois, une propriété plus riche, la différentiabilité, entraîne la continuité et la dérivabilité le long de n'importe quelle direction. La fonction $y = f(x)$ est *différentiable* au point X^0 du domaine de définition E s'il existe un vecteur $P = (p_1, ..., p_n) \in \mathbb{R}^n$ — dépendant de f et de X^0 — vérifiant l'égalité

$$f(X^0 + W) - f(X^0) = \sum_{i=1}^{n} p_i w_i + \varepsilon(W) \cdot \|W\|$$

où $\varepsilon(W)$ est une fonction réelle tendant vers 0 lorsque $\|W\| = (w_1^2 + ... + w_n^2)^{1/2} \to 0)$ pour tout vecteur

$$W = (w_1, ..., w_n) \in \mathbb{R}^n \text{ tel que } X^0 + W \in E.$$

L'expression $df = \sum_{i=1}^{n} p_i w_i$ s'appelle la différentielle de f en X^0 pour l'accroissement W. Le vecteur P s'appelle le *gradient* de f en X^0 et se note $\text{grad}_{X^0} f$.

Toute fonction différentiable en un point X^0 est continue en ce point et y admet des dérivées partielles finies selon $\left(\frac{\partial f}{\partial x_i}\right)_{X^0} = p_i$ pour tout i; par contre, la continuité en un point X^0 et l'existence de dérivées partielles finies en ce point n'entraînent pas la différentiabilité de la fonction en X^0 ; il s'agit d'un concept plus fort que l'on retrouve pour les fonctions de variable complexe avec la notion d'holomorphie. On convient de noter dx_i au lieu de w_i et la différentielle s'écrit donc :

$$df = \frac{\partial f}{\partial x_1} dx_1 + \frac{\partial f}{\partial x_2} dx_2 + ... + \frac{\partial f}{\partial x_n} dx_n$$

Considérons une fonction $z = f(x, y)$ définie et différentiable sur un ensemble $E \subseteq \mathbb{R}^2$. On peut interpréter le couple (x, y) comme un point du plan $(x, y, 0)$ de $\mathbb{R}^3$; les points $(x, y, f(x, y))$ de $\mathbb{R}^3$ forment alors la surface représentative de la fonction. L'accroissement Δz de la fonction résultant du passage de (x, y) à $(x + dx, y + dy)$ s'écarte de la différentielle par la quantité $\varepsilon(W) \cdot \|W\|$ où $\lim_{\|W\| \to 0} \varepsilon(W) = 0$, qui est donc négligeable par rapport à df (sauf lorsque $df = 0$). Ceci signifie géométriquement qu'il est — sauf exceptions — possible d'approcher la surface représentant la fonction au voisinage d'un point par le plan tangent à cette surface en ce point *(fig. 16 bis).*

Lorsque la fonction $f(X)$ admet des dérivées partielles finies en tout point $X \in E$, chacune de celles-ci est à son tour une fonction de $X = (x_1, ..., x_n)$ définie sur E. Ceci conduit à la notion de *dérivée partielle seconde;* par exemple, supposons que chaque $\frac{\partial f}{\partial x_i}$ $(i = 1, 2, ..., n)$ soit dérivable par rapport à tout x_k, on obtiendra les dérivées du second ordre :

$$\frac{\partial}{\partial x_1}\left(\frac{\partial f}{\partial x_1}\right) = \frac{\partial^2 f}{\partial x_1^2}; \quad \frac{\partial}{\partial x_2}\left(\frac{\partial f}{\partial x_1}\right) = \frac{\partial^2 f}{\partial x_2\, \partial x_1};$$

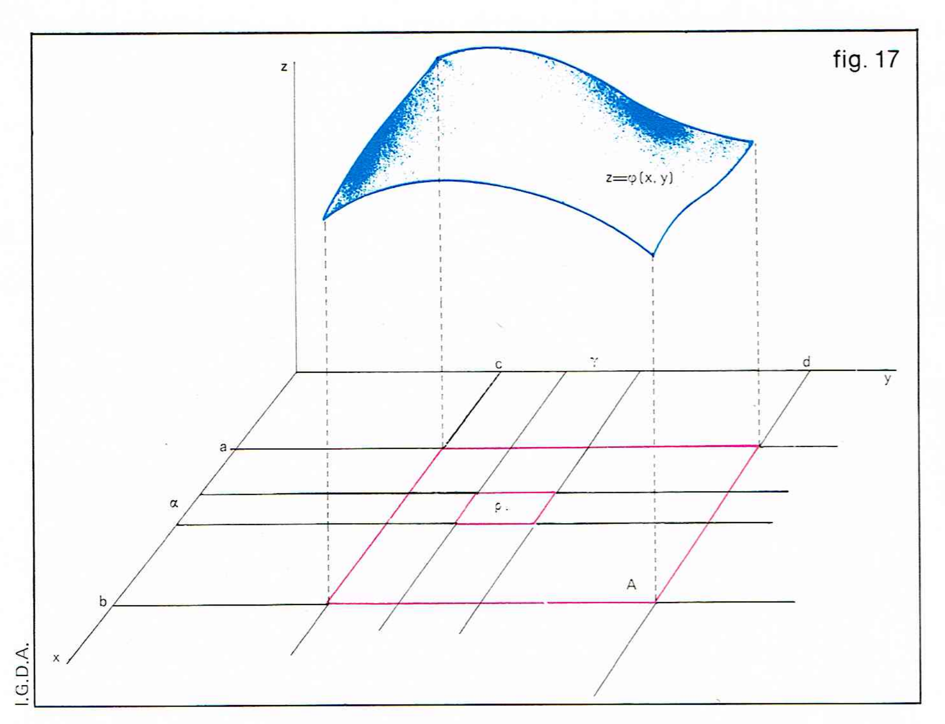
$$..., \frac{\partial}{\partial x_n}\left(\frac{\partial f}{\partial x_1}\right) = \frac{\partial^2 f}{\partial x_n\, \partial x_1};$$

et aussi $\frac{\partial}{\partial x_j}\left(\frac{\partial f}{\partial x_k}\right) = \frac{\partial^2 f}{\partial x_j\, \partial x_k}$.

◄ ***Page ci-contre en haut, tableau 14 : développement en série des fonctions analytiques usuelles.***

◄ ***Figure 16 : interprétation graphique du concept d'intégrale définie : recherche de l'aire du trapèze curviligne ABDC.***

◄ ***Page ci-contre en bas, tableau 15 : recherche de primitives.***



▲ ***Figure 17 :* *décomposition d'un rectangle A en rectangles partiels de surface ρ_{ik}, pour l'intégration de la fonction $z = \varphi(x, y)$.***

On pourrait donc être tenté de conclure à l'existence de $A_n^2 = n(n-1)$ dérivées du second ordre.

Un résultat classique connu sous le nom de **théorème de Schwartz** montre que, si $\dfrac{\partial^2 f}{\partial x_i \, \partial x_j}$ et $\dfrac{\partial^2 f}{\partial x_j \, \partial x_i}$ existent et sont finies dans un voisinage de X^0 et sont continues en X^0, alors

$$\left(\frac{\partial^2 f}{\partial x_i \, \partial x_j}\right)_{X^0} = \left(\frac{\partial^2 f}{\partial x_j \, \partial x_i}\right)_{X^0}$$

De façon analogue, on définirait les *dérivées du troisième ordre*, par exemple $\dfrac{\partial^3 f}{\partial x_1 \, \partial x_2 \, \partial x_3}$, $\dfrac{\partial^3 f}{\partial x_2 \, \partial x_1^2}$, etc., ou *d'ordre encore supérieur*, auxquelles on peut étendre le théorème d'interversion de l'ordre des dérivations sous les hypothèses vues ci-dessus ; par exemple

$$\frac{\partial^3 f}{\partial x_1^2 \, \partial x_2} = \frac{\partial^3 f}{\partial x_2 \, \partial x_1^2}$$

De même, on définira la *différentielle seconde, troisième, ..., n-ième* de f ; par exemple, pour une fonction $f(x, y)$ admettant des dérivées secondes continues :

$$d^2 f = \frac{\partial^2 f}{\partial x^2}(dx)^2 + 2\frac{\partial^2 f}{\partial x \, \partial y} dx \, dy + \frac{\partial^2 f}{\partial y^2}(dy)^2.$$

Il est alors possible de généraliser l'approximation de l'accroissement de f par sa différentielle en montrant, sous des hypothèses de régularité pour la fonction, la formule de Taylor à plusieurs variables, à l'ordre k :

$$f(X^0 + W) - f(X^0) = df + \frac{1}{2!} d^2 f + \ldots$$

$$+ \frac{1}{(k-1)!} d^{k-1} f + \frac{1}{k!} (d^k f)_{X^0 + \theta W}, \text{ où } \theta \in]0, 1[$$

(il est à noter que chacune des expressions $d^m f$ est en fait un polynôme homogène de degré m par rapport aux variables $w_1, w_2, \ldots, w_n$).

Dans de nombreux problèmes existe un lien de dépendance entre variables d'une même fonction ; par exemple, si l'on se donne deux variables x et y et la fonction $\varphi(x, y) = 8y^3 + 2xy^2 - x^2 y - x^3$, il suffit d'assujettir $\varphi(x, y)$ à être nul pour créer un tel lien entre x et y, que l'on dit être implicite. Géométriquement, ceci conduit à dire que l'on ne considérera pas la surface $z = \varphi(x, y)$ mais la courbe obtenue sur le plan $z = 0$ par intersection ; on peut donc se demander s'il existe une correspondance qui, à tout x choisi dans un certain intervalle I, associe un unique $y = y(x)$ tel que $\varphi(x, y(x)) = 0$ pour tout $x \in I$. Ainsi, dans notre exemple, on peut voir que les couples (x, y) tels que $y = \dfrac{x}{2}$ vérifient $\varphi(x, y) = 0$.

Sans chercher à donner une forme *explicite* de cette relation, le théorème suivant précise la fonction implicite par ses propriétés : soit $\varphi(x, y)$ définie sur le rectangle $x_0 - a \leqslant x \leqslant x_0 + a$, $y_0 - b \leqslant y \leqslant y_0 + b$, continue sur ce rectangle et y admettant des dérivées partielles $\dfrac{\partial \varphi}{\partial x}$ et $\dfrac{\partial \varphi}{\partial y}$ continues, avec $\dfrac{\partial \varphi}{\partial y} \neq 0$ en tout point. Si

$$\varphi(x_0, y_0) = 0,$$

alors il existe un intervalle ouvert $x_0 - \alpha < x < x_0 + \alpha$, avec $\alpha \leqslant a$, et une fonction *unique* $y = f(x)$ définie sur cet intervalle telle que $y_0 = f(x_0)$, $\varphi(x, f(x)) = 0$ pour tout $x \in]x_0 - \alpha, x_0 + \alpha[$, qui soit continue et dérivable ; la dérivée de cette fonction est alors donnée par :

$$f'(x) = -\frac{\dfrac{\partial \varphi}{\partial x}}{\dfrac{\partial \varphi}{\partial y}}$$

L'*intégration* des fonctions de plusieurs variables généralise la notion d'intégrale définie $\int_a^b f(x)\, dx$ mais peut être replacée dans le cadre général de la « mesure » (voir *Intégration des fonctions*). Par exemple, soit $z = \varphi(x, y)$ une fonction de deux variables définies sur le rectangle $A = [a, b] \times [c, d]$; en partageant le segment $[a, b]$ à l'aide des points α_i et le segment $[c, d]$ à l'aide des points γ_k, on découpe le rectangle R en petits rectangles de surface ρ_{ik}. On peut alors considérer (*fig.* 17) les parallélépipèdes dont les bases sont ces petits rectangles, et limités par la surface représentant $z = \varphi(x, y)$, soit au point le plus bas, soit au point le plus haut ; la somme de leurs volumes, lorsqu'elle a une limite pour une surface ρ_{ik} de plus en plus petite, définira alors, par cette limite, la valeur de l'intégrale de $\varphi(x, y)$ sur le domaine A : $\iint_A \varphi(x, y)\, dxdy$. Ceci peut évidemment se généraliser à un domaine A quelconque, mais mesurable (au sens de Jordan). Dans le même ordre d'idée, on peut définir de façon analogue les intégrales triples

$$\iiint_V \varphi(x, y, z)\, dxdydz,$$

ou multiples en général

$$\iint \ldots \int_E \varphi(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \ldots dx_n.$$

Le calcul des *intégrales doubles* (ce qui se généralise avec *intégrales multiples*) se fait par le **théorème de Fubini** qui permet de se ramener à celui de deux intégrales relatives chacune à une seule variable. Ainsi, dans le cas d'un domaine rectangle $A = [a, b] \times [c, d]$, si $\iint_A \varphi(x, y)\, dxdy$ existe, on a :

$$\iint_A \varphi(x, y)\, dxdy = \int_c^d dy \int_a^b \varphi(x, y)\, dx = \int_a^b dx \int_c^d \varphi(x, y)\, dy$$

On considère donc d'abord φ comme fonction de x, on l'intègre sur $[a, b]$ et le résultat obtenu qui dépend de y, soit $\psi(y)$, est intégré sur $[c, d]$ (la seconde égalité procède en sens inverse). Dans le cas d'un domaine C quelconque, on établit un résultat analogue :

$$\iint_C \varphi(x, y)\, dxdy = \int_c^d dy \int_{\alpha(y)}^{\beta(y)} \varphi(x, y)\, dx = \int_a^b dx \int_{\gamma(x)}^{\delta(x)} \varphi(x, y)\, dy,$$

où les intervalles d'intégration ne sont plus fixes mais dépendent de la frontière de C (*fig.* 18).

Ainsi, le calcul de $\iint_C (x+y)\,dxdy$, où C est la partie du premier quadrant comprise entre les courbes $x \mapsto e^x$, $x \mapsto e^{-x}$ et la droite $x = 2$ (*fig.* 19), donne :

$$\iint_C (x+y)\,dxdy = \int_0^2 dx \int_{e^{-x}}^{e^x} (x+y)\,dy =$$

$$\int_0^2 \left(xe^x + \frac{1}{2}e^{2x} - xe^{-x} - \frac{1}{2}e^{-2x}\right) dx =$$

$$e^2 + 3\,e^{-2} + \frac{1}{4}(e^4 + e^{-4}) - \frac{1}{2} \# 20{,}96.$$

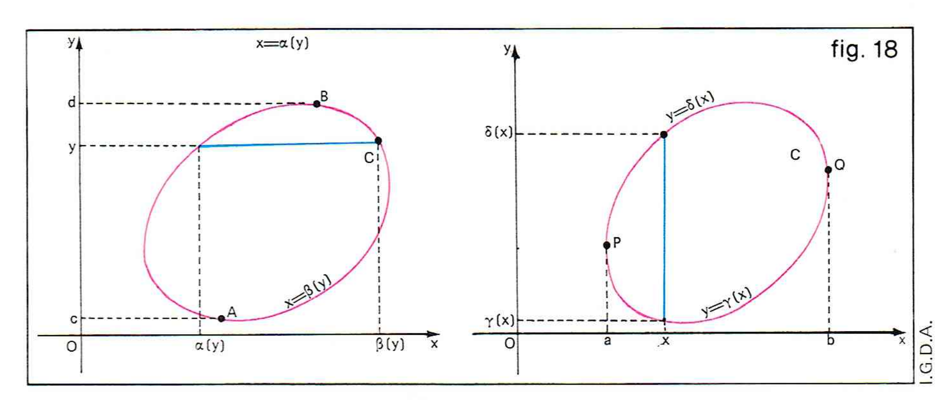
Le calcul intégral se généralise encore avec l'introduction des intégrales de surface, de volume, etc., ou intégrales vectorielles, que l'on traite plus aisément par l'outil des formes différentielles, aspect particulièrement riche du calcul différentiel sur les espaces de Banach où la notion de dérivée est remplacée par celle d'application linéaire tangente : on peut ainsi traiter l'ensemble des problèmes — quel que soit le nombre de variables — dans un cadre global.

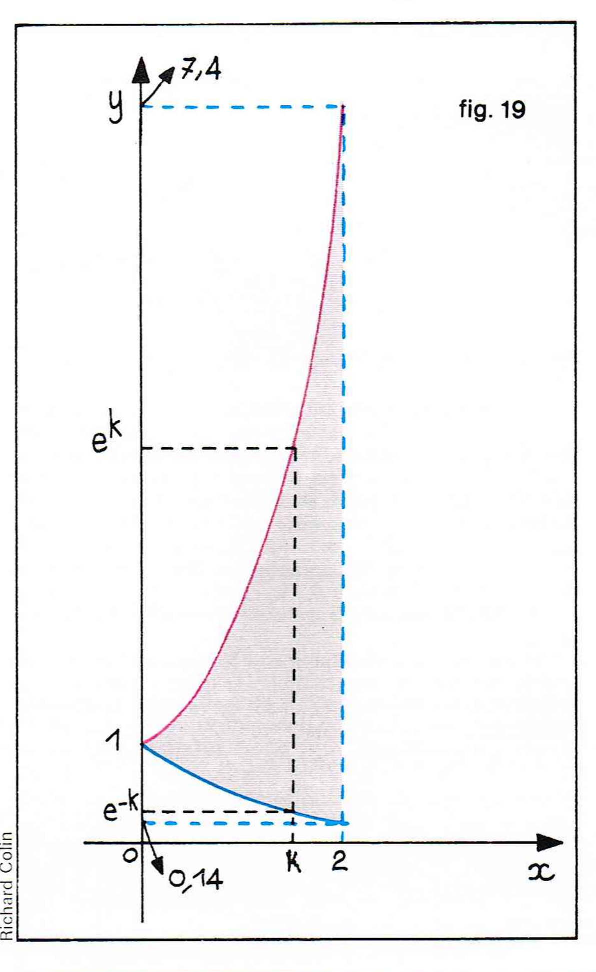
Intégration des fonctions

La mesure

Certainement l'un des aspects les plus connus de l'analyse mathématique, la *théorie de l'intégration* a connu avec la *théorie de la mesure* (et ses prolongements en probabilités) un développement exceptionnel. Une première impression pourrait la faire paraître bien théorique, et surtout bien éloignée des soucis des mathématiciens de l'Antiquité, par exemple. On peut essayer de remédier à cette apparence, car, bien qu'il s'agisse en fait d'une théorie délicate, et dont les méthodes sont difficiles à utiliser, il est possible de mettre en valeur les concepts de base et les raisons qui ont poussé aux diverses généralisations successives que l'on peut recenser depuis Eudoxe (vers 406-vers 355 avant J.-C.) jusqu'à H. Lebesgue (1875-1941), ou même Daniell (1889-1946).

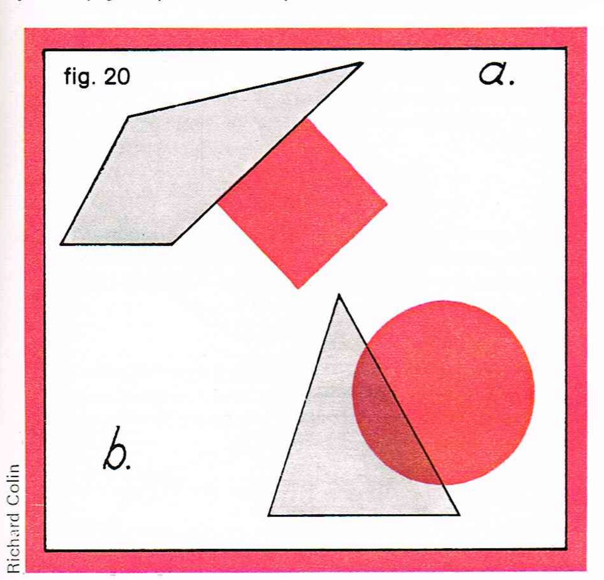
Le but poursuivi est, comme son nom l'indique, de construire un outil théorique permettant de « *mesurer* », et c'est encore une fois un problème de géométrie que l'on trouve à la source de cette démarche. Les formules donnant la surface d'un triangle, d'un quadrilatère, etc., constituent le support théorique de la mesure (d'arpentage, par exemple). Mais l'extension de nos connaissances entraîne, bien entendu, à chercher des possibilités de calcul pour des surfaces moins simples, planes d'abord, puis gauches. C'est donc bien un problème de *mesure d'ensembles* que l'on veut résoudre. Par conséquent, il s'agit de définir les propriétés que l'on peut juger nécessaires à une telle opération et les ensembles sur lesquels on pense effectuer cette mesure. Sur le premier point, on cherche à faire correspondre l'addition (comme pour les aires) des mesures d'ensembles à leur réunion disjointe (*fig.* 20) ; et d'autre part, on cherche à rendre cette

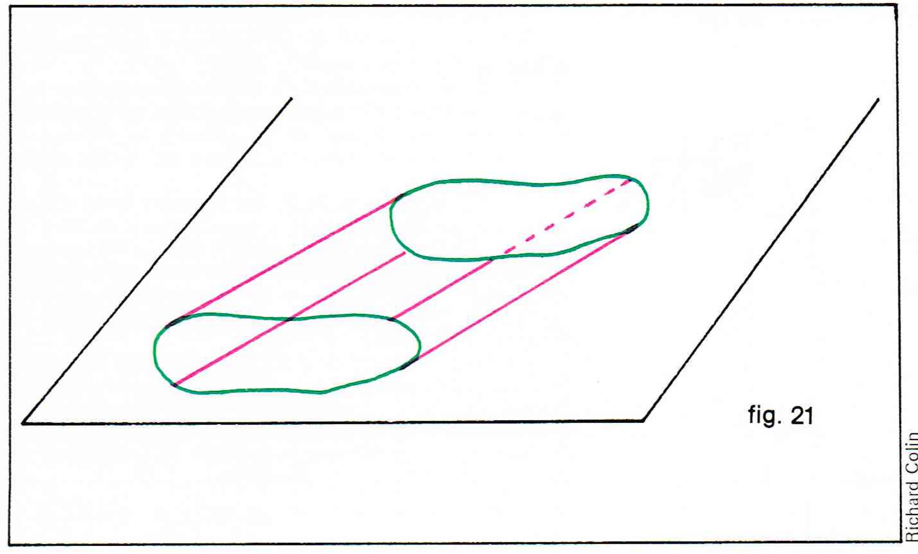


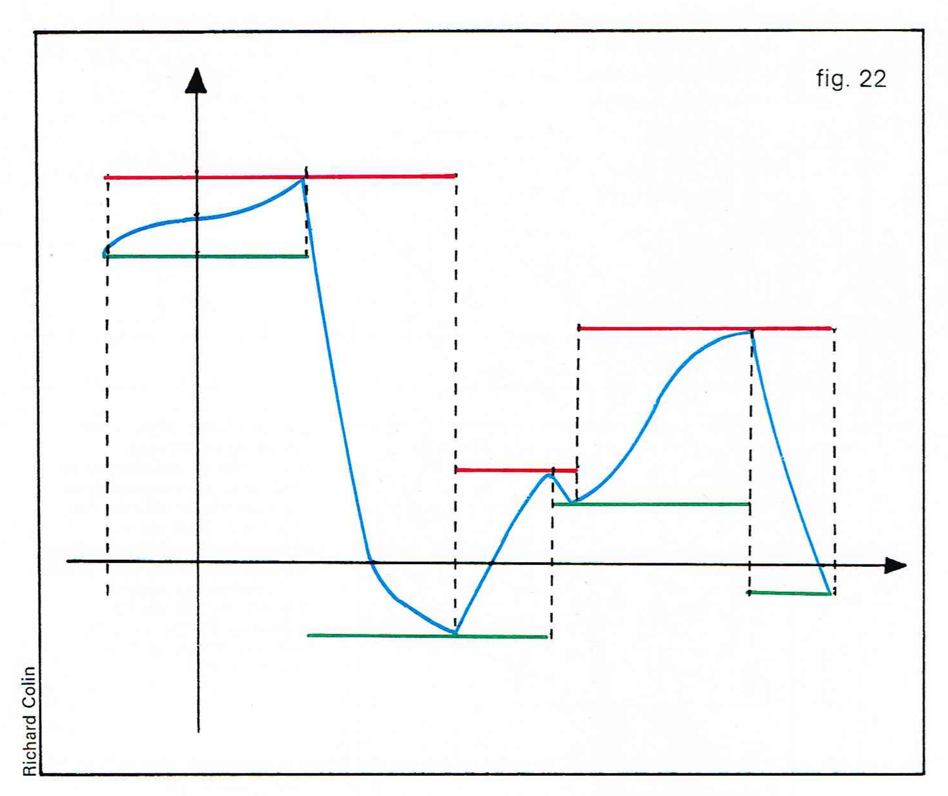


▲◄ ***En haut, figure 18 :** sur chaque dessin, la courbe C est partagée en 2 arcs correspondant à 2 fonctions distinctes, soit de x, soit de y, et qui donnent les limites d'intégration selon l'ordre d'intégration choisi.*
***A gauche, figure 19 :** domaine D* $\{(x, y) \in \mathbb{R}^2$ *tels que* $e^{-x} \leqslant y \leqslant e^x$ *et* $0 \leqslant x \leqslant 2\}$.

▼ ***A gauche, figure 20 :** a, addition des aires ; b, non-addition des aires.*
***A droite, figure 21 :** conservation de la mesure après isométrie.*







▲ ***Figure 22 : encadrement d'une fonction réelle (en bleu) par deux fonctions étagées (en rouge et en vert).***

mesure invariante après un « déplacement sans transformation », c'est-à-dire par exemple, dans le cas du plan, une isométrie, ou bijection du plan pour laquelle les distances entre points sont conservées (*fig.* 21). On est donc amené à définir, à partir de l'ensemble X sur lequel on désire mesurer, une classe de sous-ensembles, $\mathcal{A}$ (ceux que l'on mesurera), et une application $m : \mathcal{A} \to \mathbb{C}$ donnant la mesure. De nombreuses difficultés ont surgi dès que l'on a pris pour $\mathcal{A}$ l'ensemble $\mathfrak{P}$ (X) des parties de X (problème de la mesure universelle, résolu par Banach).

C'est pourquoi, pour établir une cohérence avec les propriétés envisagées pour l'application m, on choisit une *classe stable* pour les opérations de réunion et de différence :

$$A \in \mathcal{A} \quad \text{et} \quad B \in \mathcal{A} \Rightarrow A \cup B \in \mathcal{A}$$
$$\text{et} \quad A - B = \complement_A (A \cap B) \in \mathcal{A}$$

(il y a donc aussi stabilité pour l'intersection) : une telle classe $\mathcal{A}$ est appelée *algèbre de Boole* ou *clan*. On dira alors qu'un espace X est mesuré lorsqu'on a défini sur cet ensemble un clan $\mathcal{A}$ et une application (dite mesure) $m : \mathcal{A} \to \mathbb{R}^+$ vérifiant :

$$\begin{cases} m\,(A \cup B) = m\,(A) + m\,(B) - m\,(A \cap B) \\ m\,(\varnothing) = 0 \end{cases}$$

On le note alors $(X, \mathcal{A}, m)$; en voici deux exemples particulièrement importants :

— Dans l'ensemble X, la famille $\mathcal{A}$ des parties de X ayant un nombre fini d'éléments est un clan, et l'application $m : A \mapsto n_A$ qui, à tout $A \in \mathcal{A}$, associe le nombre n_A d'éléments de A, est alors une mesure sur X, dite *mesure naturelle*.

— Sur la droite réelle $\mathbb{R}$, les réunions finies d'intervalles bornés disjoints deux à deux forment un clan, dit *clan de Lebesgue ;* on définit alors la *mesure de Lebesgue sur* $\mathbb{R}$ en posant pour tout intervalle borné I d'extrémités a et b, $m\,(I) = b - a$, $(b \geqslant a)$. La mesure de tout ensemble du clan de Lebesgue est alors déduite par additivité. On note dx cette mesure.

La mesure — ainsi définie — des ensembles trouve son prolongement naturel dans la mesure des fonctions étagées par l'intermédiaire des *fonctions caractéristiques* d'ensembles (on dit parfois *indicatrices*) ; si $(X, \mathcal{A}, m)$ désigne un espace mesuré, la *fonction caractéristique* φ_A d'un ensemble $A \in \mathcal{A}$ est définie par :

$$\varphi_A\,(x) = 1 \quad \text{si } x \in A \quad \text{et} \quad \varphi_A\,(x) = 0 \quad \text{si } x \in \complement_X A.$$

C'est la structure d'espace vectoriel de l'ensemble $\mathfrak{I}$ des fonctions φ_A qui amène alors à formaliser par des notions linéaires l'intégrale de Riemann.

L'intégrale de Riemann

L'idée de base de l'intégrale au sens de Riemann est de s'assurer d'un outil mesurant les ensembles définis par des courbes représentatives de fonctions, et respectant la règle d'additivité. Pour cela, on part du cas le plus simple d'une fonction caractéristique d'ensemble, puis on généralise à une combinaison linéaire de ces fonctions, puis à des fonctions « approchées » par de telles combinaisons.

Dans un espace mesuré $(X, \mathcal{A}, m)$, à tout ensemble $A \in \mathcal{A}$, on associe sa fonction caractéristique φ_A et l'on pose : $I\,(\varphi_A) = \varphi \int_A dm = m\,(A)$. Par la linéarité de cette correspondance, on peut définir l'intégrale (au sens de Riemann) d'une fonction étagée :

$$f = \sum_{i=1}^{i=n} \alpha_i \, \varphi_{A_i}$$

par :

$$I\,(f) = \int f\,dm = \sum_{i=1}^{i=n} \alpha_i m\,(A_i) \quad \text{avec } \alpha_i \in \mathbb{R}.$$

Il s'agit alors d'étendre cette forme linéaire à un ensemble de fonctions plus vaste que l'ensemble ε des fonctions étagées : les fonctions nulles en dehors d'un ensemble A du clan $\mathcal{A}$ et bornées. Ceci est possible de la façon suivante : une telle fonction φ peut être minorée et majorée respectivement par deux fonctions étagées : $f < \varphi < g$ (*fig.* 22), et il est clair que $I\,(f) \leqslant I\,(g)$. Par conséquent, les nombres

$$\int_* \varphi\,dm = \sup \left\{ \int f\,dm \text{ pour } f \in \varepsilon,\ f < \varphi \right\}$$

$$\text{et } \int^* \varphi\,dm = \inf \left\{ \int g\,dm \text{ pour } g \in \varepsilon,\ g > \varphi \right\}$$

sont liés par la relation : $\int_* \varphi\,dm \leqslant \int^* \varphi\,dm.$

On désigne par $\tilde{\varepsilon}$ l'ensemble des fonctions φ pour lesquelles il y a égalité ; c'est un espace vectoriel — pour les lois classiques d'addition et de produit par un scalaire — et on a donc étendu la forme linéaire positive I définie sur ε en une forme linéaire positive $\tilde{I}$ définie sur $\tilde{\varepsilon}$ dès que l'on pose :

$$\tilde{I}\,(\varphi) = \int_* \varphi\,dm = \int^* \varphi\,dm$$

Il s'agit là d'une démarche analogue au « *principe d'exhaustion* » d'Eudoxe de Cnide (v. 406-v. 355 av. J.-C.) par lequel on essayait d'approcher l'aire d'un cercle par les surfaces des polygones réguliers inscrits dont on double successivement le nombre de côtés (*fig.* 23) *;* les « *sommes de Darboux* » en sont une autre expression, moins claire toutefois (*fig.* 24).

Sur l'espace $X = [a, b] \in \mathbb{R}$ où l'on prend le clan $\mathcal{A}$ engendré par les intervalles $[a, d[$ et $[c, b]$, une fonction définie croissante sur $[a, b]$, φ, permet de définir une mesure :

$$\begin{cases} m\,([a, x[) = \varphi\,(x) \\ m\,([a, b]) = \varphi\,(b) \end{cases} \quad \text{pour } x \in \,]a, b[$$

par rapport à laquelle toutes les fonctions continues réelles sont intégrables. L'intégrale d'une fonction g selon cette mesure est dite *intégrale de Riemann-Stieltjès* et se note :

$$\int_a^b g\,d\varphi.$$

Ceci permet de définir l'intégrale d'une fonction continue réelle g par rapport à une fonction à variation bornée ψ puisque celle-ci est différence de deux fonctions croissantes φ_1 et φ_2 :

$$\int_a^b g\,d\psi = \int_a^b g\,d\varphi_1 - \int_a^b g\,d\varphi_2 \quad \text{si} \quad \psi = \varphi_1 - \varphi_2$$

On peut noter que l'on retrouve l'intégrale classique en prenant $\varphi : x \mapsto x$.

Cette notion joue un rôle très particulier : on montre facilement que l'intégrale de Riemann-Stieltjès est une forme linéaire continue sur l'espace de Banach $\mathbb{R}$ [a, b] (muni de la topologie de la convergence uniforme), mais de plus, un très célèbre — et délicat — théorème, le **théorème de F. Riesz,** montre que : *toute forme linéaire continue sur cet espace est une intégrale de ce type.* C'est cette notion que l'on utilise pour définir les intégrales curvilignes.

L'intégrale au sens de Lebesgue

On peut envisager sous deux aspects l'apport de Lebesgue à la théorie de l'intégration. Tout d'abord, au sens « fonctionnel », de même que les nombres réels figurent une extension des nombres rationnels (construction de $\mathbb{R}$ par les coupures ou bien les suites de Cauchy), les fonctions intégrables au sens de Lebesgue étendent, elles, les fonctions intégrables au sens de Riemann (et plus particulièrement les fonctions continues, nulles en dehors d'un ensemble compact). D'autre part, d'un point de vue « ensembliste » et plus rapporté à l'idée générale de mesure, les développements de Lebesgue ont permis de s'affranchir de la simple additivité, et de définir la mesure pour une plus vaste classe d'ensembles, grâce à *l'idée d'additivité dénombrable* posée par Émile Borel (1871-1956) qui se traduit par l'hypothèse suivante : « la réunion d'une famille dénombrable d'éléments de $\mathcal{A}$ est encore un élément de $\mathcal{A}$ », qui donne le nom de *tribu* à une famille $\mathcal{A}$ de parties d'un ensemble X formant déjà un clan. La cohérence avec la notion de mesure est ensuite assurée par la définition d'une mesure *dénombrablement additive* (ou bien *σ-additive*), c'est-à-dire telle que, pour toute famille dénombrable $(A_n)_{n \in \mathbb{N}}$ d'éléments de $\mathcal{A}$ deux à deux disjoints ($A_p \cap A_q = \varnothing$ dès que $p \neq q$), on ait :

$$m\left(\bigcup_{n \in \mathbb{N}} A_n\right) = \sum_{n \in \mathbb{N}} m(A_n).$$

Le premier membre a bien un sens puisque l'on suppose que $\mathcal{A}$ est une tribu, le second membre est donc un nombre réel positif ou $+\infty$. Au sens de Lebesgue, un espace mesuré sera donc un *triplet* (X, $\mathcal{A}$, m) où $\mathcal{A}$ est une tribu et m une mesure σ-additive.

C'est avec les *mesures de Radon* que se dégage le mieux l'aspect linéaire de l'intégration; elles seront définies sur des espaces X supposés localement compacts (voir *Topologie*). On désigne par $\mathcal{K}$ (X) l'espace vectoriel des fonctions continues sur X, nulles hors d'un compact de X.

Une *mesure de Radon positive* sur X est une application $m : \mathcal{K}(X) \to \mathbb{C}$, linéaire

$$\text{donc } m(\alpha f + \beta g) = \alpha m(f) + \beta m(g)$$

et positive

$$\text{donc } m(f) \geqslant 0 \quad \text{si } f \geqslant 0.$$



◄ ***Le mathématicien français* Gaston Darboux *(1842-1917) qui a réalisé de nombreux travaux sur les courbes et surfaces algébriques et sur les applications géométriques du calcul infinitésimal.***

Palais de la Découverte - Paris

Outre la mesure de Lebesgue, mentionnée plus haut, on peut citer :

— la *mesure de Dirac* en un point a, notée δ_a, qui est la forme linéaire $f \mapsto f(a)$;

— *la mesure de densité p* (où p est une fonction réelle, continue, positive) par rapport à la mesure de Lebesgue :

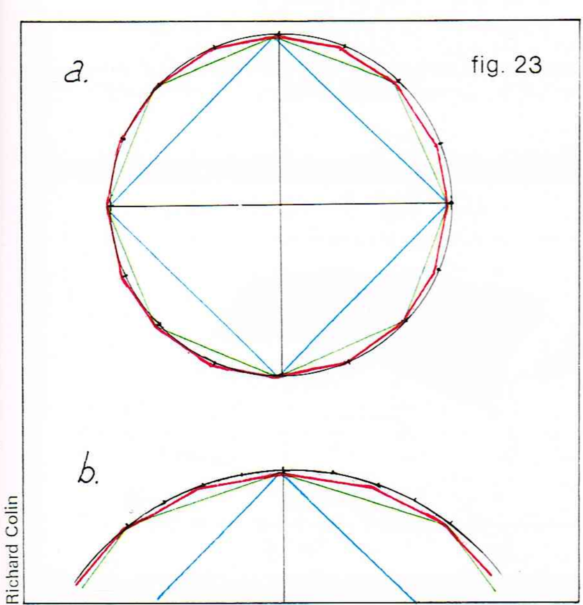
$$f \mapsto \int f(x)\, p(x)\, dx;$$

— la *mesure naturelle* sur un espace discret :

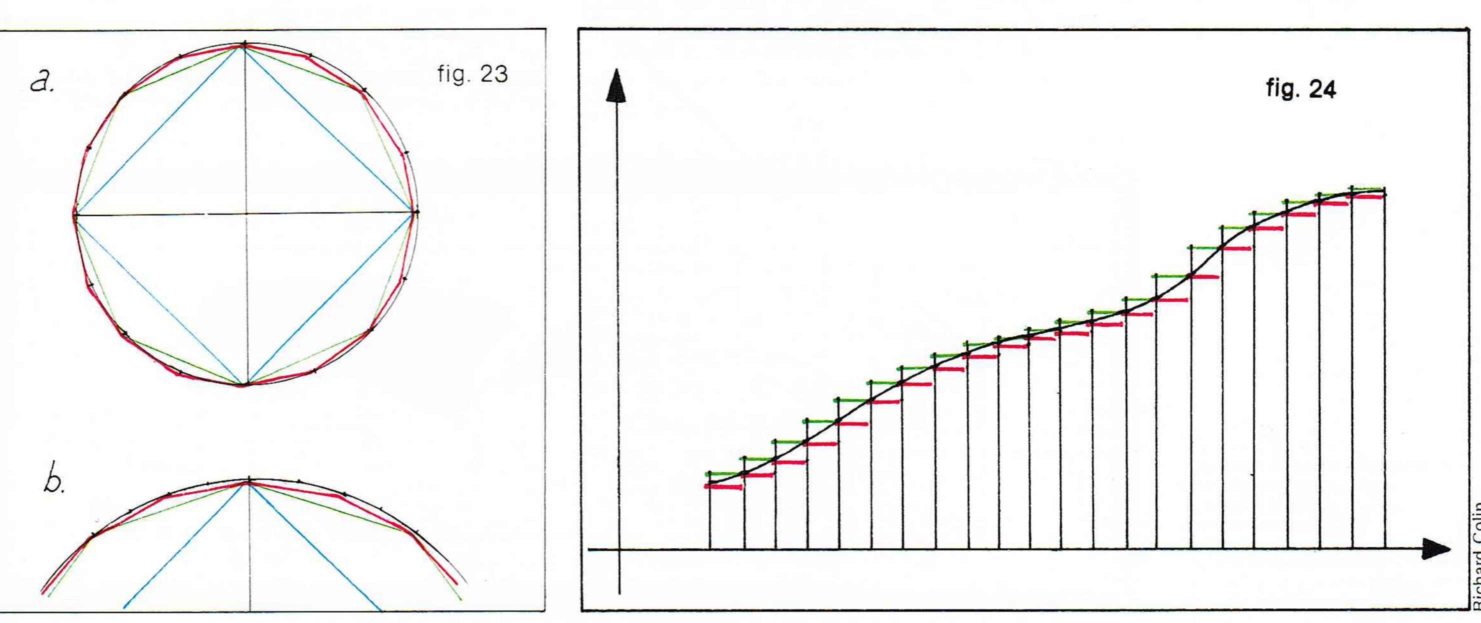
$$f \mapsto \sum_{x \in X} f(x).$$

On dit qu'une fonction f définie sur X est *intégrable pour la mesure m* (ou *m-intégrable*) si l'on peut trouver une suite $\{f_n\}_{n \in \mathbb{N}}$ de fonctions de $\mathcal{K}$ (X), telle que l'expression sup. $\int \varphi\, dm$ pour $\varphi \in \mathcal{K}(X)$ et $\varphi \leqslant |f - f_n|$ ait une limite nulle lorsque $n \to \infty$; la suite $\{f_n\}_{n \in \mathbb{N}}$ est appelée suite *d'approximation en moyenne* de f. Ceci revient à dire que l'on peut approximer la fonction f par une suite de fonctions « suffisamment régulières », puisque l'on cherche à ajuster la différence $|f - f_n|$ au plus près par

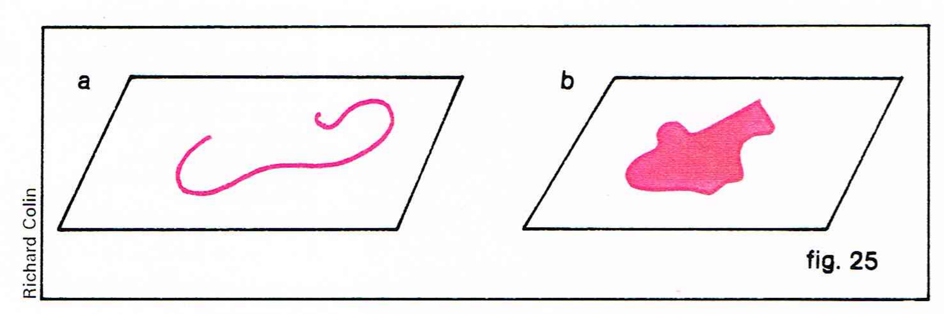
▼ ***A gauche, figure 23 : a, approximation de la circonférence ou de l'aire d'un cercle par des polygones réguliers dont le nombre de côtés est systématiquement doublé; b, détail de cette approximation. A droite, figure 24 : approximation par la méthode des sommes de Darboux de l'aire d'une surface définie par un arc de courbe.***



Richard Colin



Richard Colin



▲ *Figure 25.*
a, exemple de surface nulle : une courbe dans un plan ; b, exemple de surface non nulle : une surface dans un plan.

une fonction de moyenne (ou d'intégrale) pratiquement nulle dès que n est assez grand.

On montre que, si la suite $\{f_n\}_{n \in \mathbb{N}}$ approche f en moyenne, alors la suite des nombres $m(f_n) = \int f_n \cdot dm$ est convergente, et c'est sa limite — qui d'ailleurs reste la même quelle que soit la suite d'approximation choisie — que l'on appelle *intégrale de f par rapport à la mesure m,* soit : $m(f) = \int f \cdot dm = \lim_{n \to \infty} m(f_n) = \lim_{n \to \infty} \int f_n \cdot dm$

Espaces $\mathcal{L}^p$ et L^p

L'ensemble $\mathcal{L}^1(m)$ des fonctions m-intégrables est un espace vectoriel, puisque la mesure m est une forme linéaire ; de plus, il possède la propriété très importante :

$$f \in \mathcal{L}^1(m) \Rightarrow |f| \in \mathcal{L}^1(m).$$

Cet ensemble étend la classe des fonctions intégrables au sens de Riemann puisque :

— toute fonction $f \in \mathcal{K}(X)$ est m-intégrable (la suite d'approximation est la suite constante f pour tout n) ;

— toute fonction étagée, nulle en dehors d'un intervalle borné de $\mathbb{R}$, est intégrable pour la mesure de Lebesgue ; cette propriété est encore vraie pour une limite uniforme de fonctions étagées.

Pour la mesure de Dirac, le résultat est encore plus évident puisque toute fonction f partout finie est δ_a-intégrable ; son intégrale vaut $f(a)$.

De plus, l'application $f \mapsto \int |f| \cdot dm$ est une semi-norme sur $\mathcal{L}^1(m)$. Ce n'est pas une norme dans le cas général ; ainsi, pour la mesure de Lebesgue sur $\mathbb{R}$, tout ensemble réduit à un point (et, en raison de l'additivité dénombrable, tout ensemble dénombrable) admet une fonction caractéristique d'intégrale nulle : en effet, tout point $a \in \mathbb{R}$ peut être placé dans un intervalle ouvert $I =]a - \varepsilon, a + \varepsilon[$ où ε est arbitraire, et par conséquent $\int \varphi_{\{a\}} \cdot dx \leqslant \int \varphi_I \cdot dx = 2\varepsilon$. Il peut donc exister des fonctions $f \neq 0$ telles que $m(|f|) = 0$. Ainsi dans $\mathbb{R}^2$, la surface d'un segment de droite est nulle (*fig.* 25 et 26). Ce point fondamental est à l'origine du second point essentiel de l'intégration au sens de Lebesgue : les propriétés vérifiées *m-presque partout,* c'est-à-dire partout sauf en un ensemble de points dont la mesure est nulle. Ainsi, pour savoir si une fonction $f : X \to \mathbb{C}$ est intégrable, et pour connaître son intégrale, il suffit de connaître les valeurs prises par f en presque tout point de X. Par conséquent deux fonctions égales m-presque partout et intégrables (si l'une l'est, l'autre l'est aussi) ont même intégrale.

On est donc ainsi amené à définir sur l'espace $\mathcal{L}^1(m)$ la relation d'équivalence (R) : $f \sim g$ si et seulement si $f = g$, *m-presque partout,* puis à considérer l'espace quotient $\mathcal{L}^1 \frac{(m)}{R}$ (chaque élément de cet espace est donc la classe de toutes les fonctions de $\mathcal{L}^1(m)$, *m-presque partout* égales) que l'on désigne par $L^1(m)$ ou parfois $L^1(X, \mathcal{A}, m)$. Sur cet espace, l'application $\dot{f} \mapsto \int |f| \, dm$ ($\dot{f}$ désigne une classe selon la relation (R) dont f est un représentant) est une véritable norme ; pour la topologie induite par cette norme, $L^1(m)$ est un espace complet : c'est un *espace de Banach.* L'espace $\mathcal{L}^1(m)$ est complet pour la semi-norme $\int |f| \, dm$, mais surtout il est « approché » par l'espace $\mathcal{K}(X)$; c'est-à-dire que, pour la topologie induite par cette semi-norme, $\mathcal{K}(X)$ est dense dans $\mathcal{L}^1(m)$; ce qu'on peut encore traduire par le fait que pour toute fonction m-intégrable, il doit exister une fonction continue nulle hors d'un compact qui « approche » (en moyenne) suffisamment la première.

La notion d'intégrabilité au sens de Lebesgue généralise donc l'idée de fonction intégrable au sens de Riemann. En particulier, on peut citer deux résultats fondamentaux qui montrent bien la richesse de cette extension :

— **le théorème de Beppo-Levi ;** pour que la limite f d'une suite *croissante* de fonctions intégrables réelles $\{f_n\}_{n \in \mathbb{N}}$ soit intégrable, il faut et il suffit que la suite des nombres $\int f_n \cdot dm$ soit *majorée ;* dans ce cas, on a convergence en moyenne de f_n vers f et

$$\int f \cdot dm = \lim_{n \to \infty} \int f_n \cdot dm ;$$

— **le théorème de Lebesgue (ou de la convergence dominée) ;** si $\{f_n\}_{n \in \mathbb{N}}$ est une suite de fonctions intégrables, à valeurs complexes, *uniformément majorées* en module par une fonction intégrable, et *convergeant presque partout* vers une fonction f, alors on a convergence en moyenne, f est intégrable et

$$\int f \cdot dm = \lim_{n \to \infty} \int f_n \cdot dm.$$

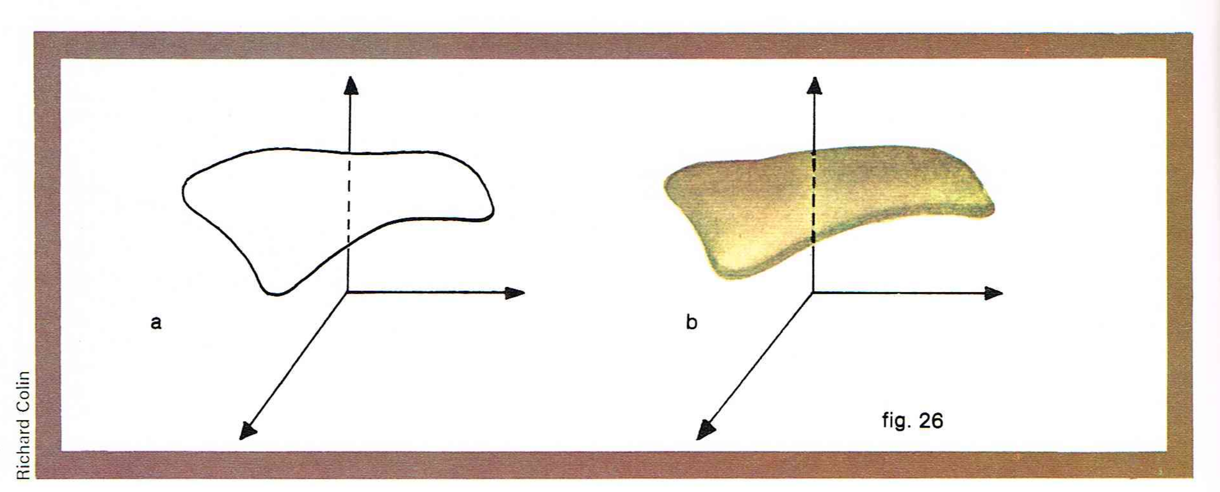
La comparaison de ces critères d'intégrabilité avec ceux connus dans le cadre de l'intégrale au sens de Riemann montre immédiatement la puissance de l'outil ainsi formé. Par exemple, une simple application du théorème de Beppo-Levi montre qu'on peut intégrer terme à terme une série de fonctions positives :

$$\int \left(\sum_n u_n\right) dm = \sum_n \left(\int u_n \cdot dm\right)$$

dès que la série des intégrales (du second membre) est convergente.

De même, le théorème de Lebesgue permet de définir des conditions affaiblies sous lesquelles il est possible de « dériver sous le signe somme ».

D'autres espaces peuvent être associés, comme $\mathcal{L}^1$ et L^1, à un espace mesuré $(X, \mathcal{A}, m)$. Ils se placent dans un cadre très général, celui des *fonctions mesurables ;* ce terme est malheureusement bien mal adapté au cadre



▶ *Figure 26.*
a, exemple de volume nul : une surface gauche dans l'espace (épaisseur nulle) ; b, exemple de volume non nul : la même forme présentant une épaisseur non nulle.

général qu'on a tracé ici, car il ne signifie absolument pas que l'on ait affaire à des fonctions que l'on puisse « mesurer »; nous désignons ces dernières — comme nous l'avons vu — par le terme de *fonctions intégrables.*

La définition de ces fonctions, que l'on pourrait juger un peu mystérieuse à ce point de notre présentation : l'application $f: X \to Y$ est m-mesurable si, quel que soit l'ouvert $\mathcal{O} \subset Y$, la fonction caractéristique de l'ensemble $K \cap f^{-1}(\mathcal{O})$ (où K est un compact quelconque de X) est m-intégrable, repose en fait sur une hypothèse « technique » nécessaire à l'exposé rigoureux de la théorie. C'est d'ailleurs une notion délicate, car, après avoir déduit quelques propriétés de ces fonctions, il semblerait que la classe des fonctions mesurables soit particulièrement vaste, au point que l'on chercherait à déterminer des fonctions qui ne le soient pas. Et c'est à l'aide d'une démonstration difficile, faisant intervenir l'axiome du choix, que l'on prouve l'existence de fonctions non mesurables. Précisons tout de suite que toute fonction *m-intégrable* est aussi *m-mesurable.*

L'espace $\mathcal{L}^p(m)$, p étant un nombre réel $1 \leqslant p < +\infty$, est défini comme l'ensemble des fonctions *m-mesurables f,* telles que $|f|^p$ soit intégrable; C' est un *espace vectoriel.* Ces espaces ne sont pas emboîtés, en général, dans un sens ou l'autre : ainsi, pour la mesure de Lebesgue sur $\mathbb{R}$, la fonction $x \mapsto |x|^{-1/2}(1+|x|)^{-1}$ pour $x \neq 0$ est dans $\mathcal{L}^1$ mais non dans $\mathcal{L}^2$, alors que la fonction $x \mapsto (1+|x|)^{-1}$ est dans $\mathcal{L}^2$ mais pas dans $\mathcal{L}^1$.

Pour chaque valeur de p $(1 \leqslant p < +\infty)$, l'espace $\mathcal{L}^p(m)$ étend l'espace $\mathcal{K}(X)$ de telle façon que $\mathcal{K}(X)$ est partout dense dans $\mathcal{L}^p(m)$: ce qui signifie donc que pour chaque fonction $\varphi \in \mathcal{K}(X)$, il existe une fonction $f \in \mathcal{L}^p(m)$ telle que le nombre

$$N_p(f-\varphi) = \left(\int |f-\varphi|^p \, dm\right)^{1/p}$$

soit rendu aussi petit que l'on veut (il y a approximation au sens de $\mathcal{L}^p$, en moyenne d'ordre p).

Intégration et recherche de primitives

A ce point d'avancement de la théorie de l'intégration, on peut se poser la question de savoir si l'on a conservé (avec l'outil ainsi mis au point) la possibilité bien connue de recherche des primitives d'une fonction, et l'intégrale indéfinie au sens usuel.

En particulier, si l'on prend $X = \mathbb{R}$, muni de la mesure de Lebesgue, que devient le résultat? L'application $\varphi : x \mapsto \int_a^x f(t)\,dt$ vérifie $\varphi'(x) = f(x)$, bien connu dès que $f \in \mathcal{K}(X)$.

La réponse est fournie par un **théorème** dû à **Lebesgue** : on désigne par $\varphi_{[a,x]}$ la fonction caractéristique de l'intervalle $[a, x] \subset \mathbb{R}$; alors l'application $x \mapsto \int f \varphi_{[a,x]}\, dm$ est *m-presque partout dérivable* et de dérivée égale à $f(x)$, dès que $f\varphi_{[a,x]}$ est m-intégrable (on dit alors que f est *localement intégrable*). On pose

$$\int f\varphi_{[a,x]}\, dm = \int_a^x f(t)\, dt$$

(on rappelle qu'ici m représente la mesure de Lebesgue sur $\mathbb{R}$), appelé *intégrale (indéfinie) de f.*

Par conséquent toute fonction localement intégrable est presque partout égale à la dérivée de son intégrale (indéfinie).

Le problème réciproque nécessite l'introduction d'un type particulier de fonctions réelles, les *fonctions absolument continues* sur un intervalle $[a, b] \subset \mathbb{R}$. Une telle fonction F satisfait à la propriété : pour tout $\varepsilon > 0$, on peut trouver $\eta > 0$ tel que, dès que les sous-intervalles $[\alpha_k, \beta_k]$ de $[a, b]$ (un nombre fini) ne se recouvrent pas et ont une longueur totale inférieure à η, alors

$$\sum_k |F(\beta_k) - F(\alpha_k)| \leqslant \varepsilon.$$

Entre autres propriétés, les fonctions absolument continues sont uniformément continues, et à variation bornée; de plus, elles peuvent s'exprimer comme différence de deux fonctions croissantes absolument continues. Mais surtout, et c'est là la réponse à la question posée, une fonction $F : [a, b] \to \mathbb{C}$ est une intégrale indéfinie si et seulement si F est absolument continue, et ce à une constante additive près :

$$F(x) = F(a) + \int_a^x F'(t)\, dt.$$

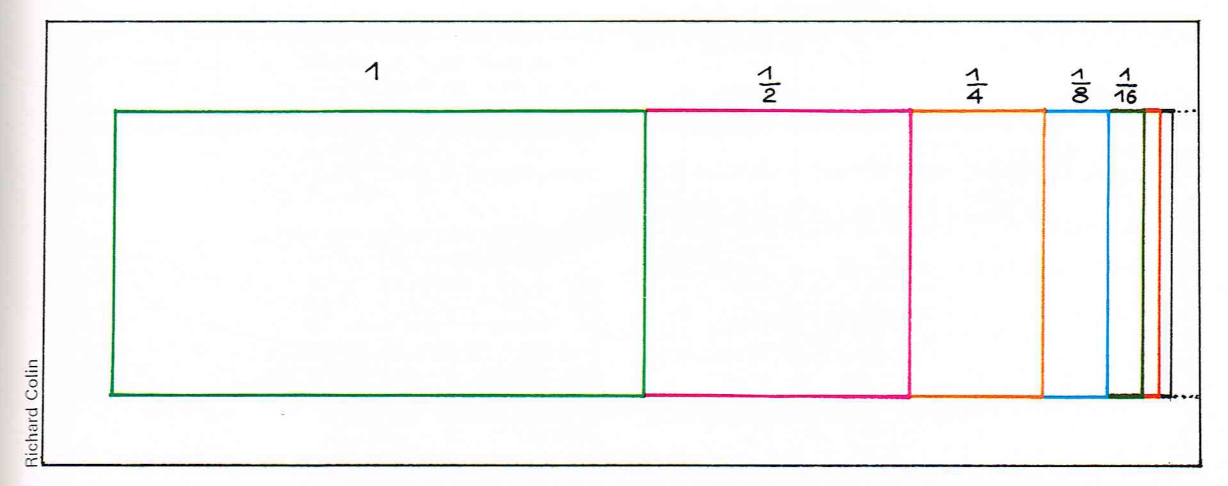
Les résultats que nous avons mentionnés ici et qui précisent donc que, relativement à la mesure de Lebesgue sur $\mathbb{R}$, l'intégration et la dérivation presque partout sont des opérations réciproques, peuvent être mis en forme dans un cadre plus général, celui des mesures de densité sur un espace X, localement compact et dénombrable à l'infini (c'est-à-dire réunion dénombrable de compacts). Si m est une mesure positive sur X et p une fonction positive, localement intégrable sur X, alors l'application $f \mapsto \int f \cdot p \cdot dm$ est une mesure de Radon positive sur X, que l'on appelle *mesure de densité p* pour la mesure m; on la note $p \cdot m$. On dit qu'une mesure positive μ sur X est *de base m* s'il existe une fonction $p \geqslant 0$ localement m-intégrable sur X, telle que $\mu = p \cdot m$.

Le **théorème de Radon-Nikodym** montre alors que, *pour que μ soit de base m, il faut et il suffit que tout ensemble m-négligeable soit μ-négligeable (on dit encore que μ est absolument continue relativement à m).*

On a là, très certainement (par l'intermédiaire des mesures de densité), la présentation la plus cohérente tant de la dérivation que de l'intégration (il suffit d'ailleurs, pour s'en convaincre, de se reporter au concept physique de densité) ; c'est ce que l'on retrouve dans les théories modernes des probabilités où l'on fait un abondant usage des formalismes et résultats de la théorie de la mesure.

La notion de fonction absolument continue permet encore de généraliser la formule d'*intégration par parties :*

$$\int_a^b f(x)\, g'(x)\, dx = [f(b)\, g(b) - f(a)\, g(a)] - \int_a^b f'(x)\, g(x)\, dx$$



◀ *Un exemple de série convergente : la série* $1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \dots + \frac{1}{2^n}$ *tend vers 2 quand n tend vers l'infini.*

— bien connue pour des fonctions de classe C^1 sur $[a, b]$ — lorsque f et g sont absolument continues sur $[a, b]$, toujours grâce aux résultats de Lebesgue sur les fonctions presque partout dérivables.

Signalons enfin que la théorie de l'intégration par rapport à une mesure de Radon positive englobe les problèmes de l'*intégration multiple* et retrouve tant le théorème de Fubini sur la décomposition d'une *intégrale n-uplet* sur un domaine de $\mathbb{R}^n$, que la formule du changement de variable.

Les séries

La notion de limite, fondamentale pour la science mathématique, est déjà présente (sous forme rudimentaire) chez les mathématiciens de l'Antiquité par l'intermédiaire des suites $s_1, s_2, ..., s_n, ...$, où l'élément s_n ($n = 1, 2, 3$, etc.) peut être un nombre (réel ou complexe), un vecteur, une fonction, etc. Une suite se présente souvent comme une succession de résultats, pour une opération répétée un nombre illimité de fois, et ainsi reliés à un « algorithme ». Parmi ces procédés, on retiendra comme particulièrement intéressants les *séries* qui se présentent comme une généralisation de l'addition, et les *produits infinis*, généralisant la multiplication.

Séries numériques

Un premier exemple élémentaire de série est celui que l'on obtient si l'on veut représenter un nombre rationnel au moyen des nombres décimaux; par exemple, la représentation décimale du nombre rationnel 1/6 est 0,166 66..., qui « équivaut » à la série :

$$\frac{1}{10} + \frac{6}{10^2} + \frac{6}{10^3} + ... + \frac{6}{10^n} + ...$$

L'intuition suggère alors d'écrire : $\frac{1}{6} = \frac{1}{10} + \frac{6}{10^2} + ...$,

c'est-à-dire d'égaler le symbole du second membre au nombre 1/6. On observe tout de suite que l'égalisation entre un tel symbole et un nombre nous entraîne à construire sur ces nouveaux êtres mathématiques des règles de calcul cohérentes avec celles des nombres et avec leurs propriétés. Ainsi, la somme des k premiers termes $\frac{1}{10} + \frac{6}{10^2} + \frac{6}{10^3} + ... + \frac{6}{10^k}$ diffère de $\frac{1}{6}$ d'aussi peu que l'on désire, à condition de choisir k suffisamment grand.

Définissons alors une série : étant donné la suite $a_0, a_1, a_2, ..., a_n$ de nombres (ou plus généralement d'éléments appartenant à un ensemble sur lequel on a défini une opération d'addition et qui est stable pour cette opération), on appelle *série* le symbole

$$a_0 + a_1 + a_2 + ... + a_n + ...,$$

aussi écrit $\sum_{n=0}^{\infty} a_n$ (a_n est le terme général de la série).

Avec cette écriture, on entend construire la suite des sommes partielles :

$s_0 = a_0, s_1 = a_0 + a_1, s_2 = a_0 + a_1 + a_2, ...,$
$s_n = a_0 + a_1 + a_2 + ... + a_n,$

et donner ainsi à la série les caractères (convergence, divergence, etc.) de cette suite.

S'il existe un nombre fini A tel que $\lim_{n + \infty} s_n = A$, la *série* est dite *convergente*, et l'on dit que A est sa somme : par exemple, la série $1 + \frac{1}{2} + \frac{1}{2^2} + ... + \frac{1}{2^n} + ...$, alors

$$s_0 = 1, \quad s_1 = 1 + \frac{1}{2} = \frac{3}{2}$$

$$s_n = \frac{1 - \frac{1}{2^{n+1}}}{1 - \frac{1}{2}} = 2\left(1 - \frac{1}{2^{n+1}}\right) \to 2 \text{ lorsque } n \to \infty.$$

Dans le cas où $\lim_{n \to + \infty} s_n = + \infty$ ou bien $- \infty$, la *série* est dite *divergente* : par exemple, $1 + 1 + 1 + ...$; alors $s_1 = 1$, $s_2 = 2$, ..., $s_n = n$, et donc s_n croît sans limite.

Enfin, lorsque $\lim_{n \to \infty} s_n$ n'existe pas, la *série* est dite *oscillante* ou *irrégulière* (par exemple, $1 - 1 + 1 - 1 + ...$; alors $s_0 = 1$, $s_1 = 0$, $s_2 = 1$, $s^3 = 0$, etc.).

Les séries convergentes et les séries divergentes sont dites *régulières*. Notons que, lorsque l'on a $\lim_{n \to \infty} s_n = \infty$ (sans être précisément $+ \infty$ ou $- \infty$), pour connaître si la série est divergente ou bien oscillante, il est nécessaire de préciser la façon dont l'axe réel a été « compactifié » : car, lorsqu'il l'a été par l'adjonction des points $+ \infty$ et $- \infty$, alors une série dont les sommes partielles sont par exemple

$$s_0 = 0, s_1 = -1, s_2 = +2, s_3 = -3, s_4 = +4, ...,$$

n'admettant pas de limite (ni $+ \infty$, ni $- \infty$) est à cause de cela oscillante, tandis que si l'axe réel a été compactifié par l'adjonction du seul point ∞, alors la suite 0, — 1, + 2, — 3, + 4, ..., admet une limite infinie, et la série est alors divergente.

Prenons quelques exemples :

— puisque $\frac{1}{n(n+1)} = \frac{1}{n} - \frac{1}{n+1}$, on constate aisément la convergence de la *série de Mengoli* dont la somme est :

$$2 = \frac{1}{1 \cdot 2} + \frac{1}{2 \cdot 3} + \frac{1}{3 \cdot 4} + ... + \frac{1}{n(n+1)} + ...;$$

— la *série* « *géométrique* »

$$\Sigma q^n = 1 + q + q^2 + ... + q^n + ...$$

(de raison q) a un caractère dépendant de la valeur de q; en effet, puisque $s_n = 1 + q + ... + q^n = \frac{1 - q^{n+1}}{1 - q}$, on reconnaît la convergence (avec une somme égale à $\frac{1}{1-q}$) lorsque $|q| < 1$, la divergence pour $|q| > 1$ et $q = 1$, tandis que la série est oscillante si $q = 1$;

— la série « *harmonique* » $1 + \frac{1}{2} + \frac{1}{3} + ... + \frac{1}{n} + ...$ est divergente et la somme partielle $s_n = 1 + \frac{1}{2} + \frac{1}{3} + ... + \frac{1}{n}$ vérifie $s_n = \text{Log}\, n + C + \varepsilon(n)$, où C = 0,57... *(constante* dite *d'Euler-Mascheroni)* et $\varepsilon(n)$ est un correctif d'approximation qui tend vers zéro lorsque n tend vers l'infini.

Pour toute série $\sum^{\infty} a_n$ on peut considérer la série $\sum^{\infty} |a_n|$; dans le cas où cette dernière série est convergente, alors la première l'est aussi, et l'on dit qu'elle est *absolument convergente*. Toutefois, une série peut converger au sens ordinaire sans être absolument convergente; ainsi en est-il de la série harmonique alternée :

$$1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + ... + (-1)^{n+1}\frac{1}{n} + ...$$

Pour toute série convergente, le terme général a_n doit tendre vers zéro lorsque $n \to \infty$; toutefois, ici encore, il s'agit d'une condition nécessaire mais non suffisante de convergence, ainsi qu'on le vérifie par l'exemple de la série harmonique $1 + \frac{1}{2} + \frac{1}{3} + ... + \frac{1}{n} + ...$

Propriétés formelles des séries

Une série se présente comme une généralisation « naturelle » de l'opération arithmétique qu'est l'addition, à travers un processus de passage à la limite (où l'on étend un classique algorithme fini à un algorithme infini). L'addition possède de nombreuses propriétés formelles : associativité, commutativité, etc. Dans quelle mesure ces propriétés se conservent-elles (ou donnent-elles lieu à des propriétés analogues) pour les séries? A cette question, on peut répondre qu'en général il n'y a pas conservation; plus précisément en trois points :

— Lorsque l'on regroupe, par une loi quelconque, un ensemble de termes consécutifs, dans une série convergente, on obtient une série convergente et de même somme; en procédant de même dans une série divergente vers $+\infty$ (respectivement $-\infty$), on retrouve une série divergente vers $+\infty$ (resp. $-\infty$). Par conséquent, la propriété d'associativité reste valable pour les séries régulières.

— Mais, à cette exception près des séries régulières, on ne peut plus rien affirmer. Par exemple, la série $1-1+1-1+\ldots$ est oscillante, alors que la série obtenue à partir de celle-ci en regroupant les termes deux à deux, c'est-à-dire : $(1-1)+(1-1)+(1-1)+\ldots$, soit $0+0+0+\ldots$, est convergente et de somme égale à zéro.

— L'examen de la commutativité nécessite d'abord de préciser la notion de « permutation dans l'ordre des termes » (particulièrement dans le cas d'une infinité de termes) et permet alors de retrouver une notion forte de convergence. Étant donné deux séries $\Sigma\, u_n$ et $\Sigma\, v_m$, on dit que l'une s'obtient à partir de l'autre par permutation de l'ordre des termes lorsque existe une correspondance biunivoque entre les rangs des termes des deux séries, de telle façon que, pour des rangs qui se correspondent, on ait des termes égaux. Par exemple :

$$\Sigma\, u_n = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \ldots$$

$$\Sigma\, v_m = 1 + \frac{1}{3} - \frac{1}{2} + \frac{1}{5} + \frac{1}{7} - \frac{1}{4} + \ldots$$

Une série convergente, telle que toute autre série obtenue par permutation dans l'ordre des termes soit aussi convergente, est dite *inconditionnellement convergente ;* cette notion est en fait équivalente à celle de la convergence absolue. Une telle série et toutes celles obtenues par le procédé décrit ont la même somme : pour elles, on a donc conservation de la propriété de commutativité.

Opérations sur les séries convergentes

Étant donné deux séries convergentes $\Sigma\, a_n = A$, $\Sigma\, b_n = B$, on définit naturellement les opérations d'*addition* des deux séries et de *multiplication* de l'une d'elles par un nombre :

$$\Sigma\, a_n + \Sigma\, b_n = \Sigma\,(a_n + b_n)$$
$$c \cdot \Sigma\, a_n = \Sigma\, c \cdot a_n$$

les séries obtenues $\Sigma\,(a_n + b_n)$ et $\Sigma\, c \cdot a_n$ sont convergentes, et leurs sommes sont respectivement $A + B$ et $c \cdot A$.

Le *produit* de deux séries $\Sigma\, a_n$ et $\Sigma\, b_n$ conduit formellement à une série double $\sum\limits_{n,\, m} a_n b_m$, pour laquelle on peut envisager divers modes de regroupement de termes pour construire une série $\Sigma\, u_k$. On se limitera ici à citer le produit selon Cauchy, dans lequel le terme u_k est construit de la façon suivante :

$$u_k = a_0 b_k + a_1 b_{k-1} + a_2 b_{k-2} + \ldots + a_k b_0 = \sum_{l=0}^{l=k} a_l b_{k-l}.$$

Cette notion de produit est la plus naturelle pour des séries entières (se reporter au paragraphe sur les *Séries entières*), puisque, dans ce cas, la série produit obtenue est elle aussi une série entière. La convergence de la série $\Sigma\, u_k$ et la relation espérée $U = A \cdot B$ entre les sommes A et B des facteurs et la somme U de la série produit ne sont toutefois assurées qu'avec des hypothèses complémentaires sur les séries $\Sigma\, a_n$ et $\Sigma\, b_n$ (la convergence absolue de ces séries, par exemple).

Critères de convergence des séries

Pour déterminer le caractère d'une série, ce n'est en général pas à la définition de la convergence que l'on a recours, mais plutôt à l'un des multiples critères exposés ci-dessous :

• *Condition de Cauchy*. Une condition nécessaire et suffisante pour que la série $\Sigma\, a_n$ soit convergente est que, pour tout nombre $\varepsilon > 0$, on puisse déterminer un indice N (dépendant en général de ε) tel que, pour tout entier $n \geqslant N$ et tout entier $p > 0$, on ait :

$$|a_{n+1} + a_{n+2} + \ldots\ldots\, a_{n+p}| < \varepsilon.$$

• *Conditions suffisantes de convergence*. Parmi celles-ci on peut citer :

— *le critère de comparaison ;* pour deux séries à termes positifs (c'est-à-dire non négatifs), $\Sigma\, a_n$ et $\Sigma\, b_n$ vérifiant $a_n \leqslant b_n$, lorsque $\Sigma\, b_n$ converge, il en est de même pour $\Sigma\, a_n$, et lorsque $\Sigma\, a_n$ diverge, alors $\Sigma\, b_n$ diverge aussi ;

— *le critère de D'Alembert ;* étant donné la série $\Sigma\, a_n$ $(a_n > 0)$, s'il existe un nombre α, $0 < \alpha < 1$, tel que $\dfrac{a_{n+1}}{a_n} < \alpha$, alors cette série est convergente ; elle est divergente si $\dfrac{a_{n+1}}{a_n} \geqslant 1$;

— *le critère de Cauchy*; étant donné la série $\Sigma\, a_n$ $(a_n \geqslant 0)$, s'il existe un nombre β, $0 < \beta < 1$, tel que : $\sqrt[n]{a_n} < \beta$, cette série est convergente ; elle est divergente si : $\sqrt[n]{a_n} \geqslant 1$.

Pour ces trois critères, les conditions (inégalités) à remplir peuvent n'être réalisées qu'à partir d'une certaine valeur de l'indice ; c'est en effet le comportement pour $n \to \infty$ qui détermine la nature de la série. Modifier un nombre fini de termes ne peut donc pas changer le caractère d'une série, mais au plus sa somme lorsque cette série est convergente.

• Un type fréquent de série est celui des *séries alternées*, c'est-à-dire celles dont deux termes consécutifs sont de signes opposés : $u_0 - u_1 + u_2 - u_3 + u_4 - \ldots$ (où $u_n > 0$). Pour qu'une telle série soit convergente, il suffit, outre la condition générale $u_n \to 0$ lorsque $n \to \infty$, qu'elle soit à termes décroissants, c'est-à-dire que l'on ait : $u_0 \geqslant u_1 \geqslant u_2 \geqslant \ldots$

Avant de développer d'autres questions, notons que, sur l'exemple qui nous a servi de présentation, le calcul de la somme de la série peut être réalisé aisément grâce à la série géométrique. On peut en effet écrire :

$$\frac{1}{10} + \frac{6}{10^2} + \frac{6}{10^3} + \frac{6}{10^4} + \ldots =$$
$$\frac{1}{10} + \frac{6}{10^2}\left(1 + \frac{1}{10} + \frac{1}{10^2} + \ldots\right) =$$
$$\frac{1}{10} + \frac{6}{10^2}\left(\frac{1}{1 - \frac{1}{10}}\right) = \frac{1}{10} + \frac{1}{15} = \frac{1}{6}$$

ce qui correspond bien à la considération de départ.

Séries de fonctions

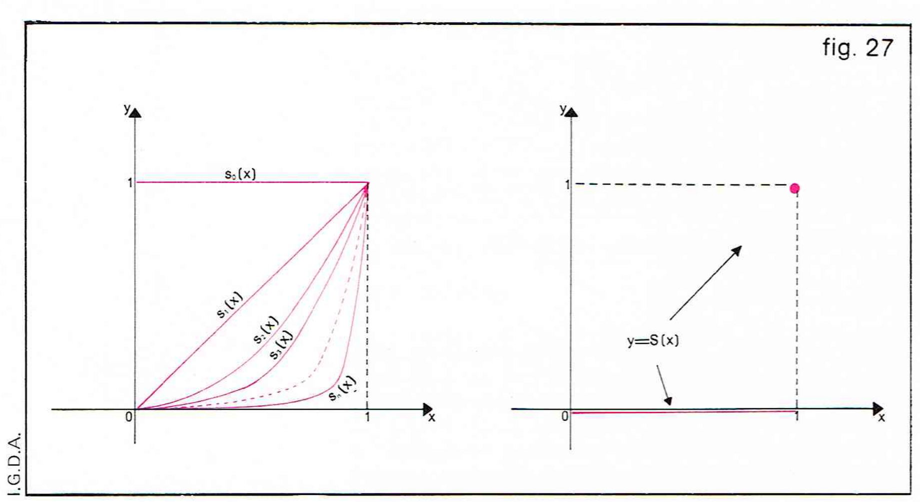
La notion introduite par les séries numériques s'étend par les séries de fonctions (plus généralement encore par les séries de fonctions vectorielles pour lesquelles on a besoin des outils de l'analyse fonctionnelle, développés au chapitre correspondant) ; dans ce paragraphe, on ne considérera que des fonctions réelles d'une seule variable réelle.

Si $\{u_n(x)\}$ $(n = 0, 1, 2, 3, \ldots)$ désigne une suite de fonctions $u_n(x)$ toutes définies sur le même ensemble $E \subset \mathbb{R}$, pour toute valeur $x_0 \in E$, $\{u_n(x_0)\}$ est une suite numérique ; le symbole $\sum\limits_{n=0}^{\infty} u_n(x_0)$ est alors défini, ainsi que toutes les notions (convergence, divergence, etc.) qui s'y rattachent. Par contre, lorsqu'on considère x comme une variable décrivant E, $\sum\limits_{n=0}^{\infty} u_n(x)$ se présente comme une série de fonctions de la variable x ; si $\sum\limits^{\infty} u_n(x)$ converge pour toute valeur $x \in E$, on dira que la série converge dans E, et la somme $S(x)$ est une fonction définie dans E.

Par exemple, la série géométrique

$$\sum x^n = 1 + x + x^2 + \ldots + x^n + \ldots$$

converge dans l'intervalle $]-1, +1[$ (intervalle ouvert) et y a pour somme $S(x) = \dfrac{1}{1-x}$. On dit que la série $\sum\limits^{\infty} u_n(x)$ est *absolument convergente* dans E si la série



▲ ***Figure 27 :** exemple d'une série de fonctions continues sur un intervalle qui converge (non uniformément) et dont la somme n'est pas continue; on l'obtient avec $u_0(x) = 1$, $u_n(x) = x^{n-1} \cdot (x-1)$ sur l'intervalle [0, 1]; les sommes partielles $s_0(x) = 1$, $s_1(x) = x$, ..., $s_n(x) = 1 + (x-1)(1 + x + ... + x^{n-1})$ montrent que $s_n(1) = 1$ pour tout $n \in \mathbb{N}$, donc que $S(1) = 1$; par contre si $x \in [0, 1[$ $s_n(x) = 1 + (x-1) \cdot \frac{x^n - 1}{x - 1} = x^n$ et donc $S(x) = 0$.*

$\sum^{\infty} |u_n(x)|$ converge dans E.

C'est en cherchant à étendre aux séries de fonctions les quelques opérations déjà connues pour des sommes finies de fonctions (par exemple, le passage à la limite, la continuité, la dérivation ou l'intégration terme à terme) que s'est dégagée la notion fondamentale de *convergence uniforme*.

Lorsque la série $\sum^{\infty} u_n(x)$ converge dans E avec une somme égale à S (x), alors, pour tout $x \in E$, si l'on fixe $\varepsilon > 0$, on peut trouver un entier $n_0 = n_0(\varepsilon, x)$ tel que, dès que $n \geqslant n_0$, on ait :

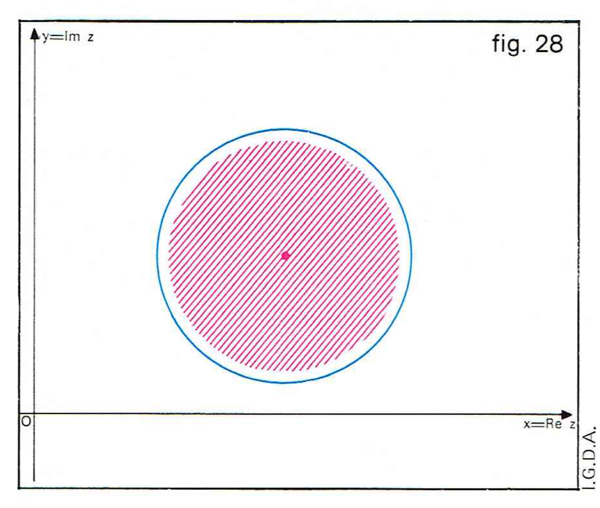
$$\left| S(x) - \sum_{k=0}^{n} u_k(x) \right| < \varepsilon.$$

Il arrive que, pour $\varepsilon > 0$ fixé, il soit possible de déterminer l'entier n_0 en ne tenant compte que de ε (par conséquent, cet entier reste le même pour toutes les valeurs de $x \in E$), de telle façon que la condition précédente soit vérifiée. C'est dans ce cas que l'on dit que la série est uniformément convergente dans E; il est évident que la notion de convergence uniforme est relative au comportement de la série sur un ensemble, et non pas en un seul point.

Pour la convergence uniforme, on a encore la **condition générale** suivante (dite **de Cauchy**) : une condition nécessaire et suffisante pour que la série $\sum^{\infty} u_n(x)$ converge uniformément dans E est que, pour tout nombre $\varepsilon > 0$,on puisse déterminer un entier $n_0 = n_0(\varepsilon)$, tel que, dès que $n \geqslant n_0$ et $p > 0$ (n et p entiers), on ait :

$$|u_{n+1}(x) + u_{n+2}(x) + u_{n+3}(x) + ... + u_{n+p}(x)| < \varepsilon$$

ce, quel que soit $x \in E$.



▶ ***Figure 28 :** cercle de convergence de la série entière $\sum_{n=0}^{\infty} c_n (z-a)^n$, z étant une variable complexe. Le cercle hachuré, concentrique et de rayon inférieur, représente un ensemble de points dans lequel la série donnée est uniformément convergente.*

Dans les applications courantes, on utilise souvent le critère suivant, qui est une condition suffisante : la série $\sum^{\infty} u_n(x)$ est uniformément convergente dans E s'il existe une série numérique $\sum^{\infty} a_n$, à termes positifs, convergente et telle que l'inégalité $|u_n(x)| \leqslant a_n$ soit vérifiée pour tout $x \in E$.

Ainsi, la série de terme général $\frac{\sin x}{2^n} = u_n(x)$ est uniformément convergente dans $\mathbb{R}$ tout entier, puisque $|\sin x| \leqslant 1$, donc que $|u_n(x)| \leqslant \frac{1}{2^n}$, cette dernière série convergeant en raison du critère de Cauchy énoncé plus haut.

D'un **théorème** très classique (dit **de la double limite**) concernant les suites de fonctions, on déduit — entre autres — le résultat suivant : la somme d'une série de fonctions continues sur l'ensemble fermé E, uniformément convergente dans E, est une fonction continue sur E.

De même, les théorèmes suivants mettent en valeur l'importance de la notion de convergence uniforme :

— **intégration terme à terme ;** si, pour toutes les valeurs entières de k ($k = 0, 1, 2, ...$), $u_k(x)$ est une fonction réelle définie sur l'intervalle fermé $[a, b]$, et intégrable sur le même intervalle, si la série $\sum^{\infty} u_k(x)$ est uniformément convergente dans $[a, b]$ et de somme $f(x)$, alors l'intégrale $\int_a^b f(x)\, dx$ existe, la série

$$\sum_{k=0}^{\infty} \left\{ \int_a^b u_k(x)\, dx \right\}$$

est convergente, et l'on a l'égalité (*fig.* 27) :

$$\int_a^b f(x)\, dx = \sum_{k=0}^{\infty} \left\{ \int_a^b u_k(x)\, dx \right\}$$

— **dérivation terme à terme ;** si la série $\sum^{\infty} u_k(x)$ converge dans l'intervalle fermé $[a, b]$, sa somme étant $f(x)$, si, pour toutes les valeurs entières de k ($k = 0, 1, 2, ...$), la fonction $u_k(x)$ est dérivable, et si la série $\sum^{\infty} u'_k(x)$ des dérivées est uniformément convergente dans le même intervalle, alors :

la fonction $f(x)$ est dérivable sur l'intervalle ouvert $]a, b[$

et on a l'égalité : $f'(x) = \sum_{k=0}^{\infty} u'_k(x)$.

Séries entières

Un cas particulier parmi les séries de fonctions, et très utilisé, est constitué par les *séries entières* ou *séries de puissances*. On appelle série entière de la variable x, centrée en a, toute série de la forme :

$$\sum_{n=0}^{\infty} c_n (x-a)^n = c_0 + c_1(x-a) + c_2(x-a)^2 + ... + c_n(x-a)^n + ...$$

où c_n ($n = 0, 1, 2, ...$), a et x sont, de la façon la plus générale, des nombres complexes. En particulier, lorsque $a = 0$ (série centrée à l'origine), la série entière prend la forme $\sum_{n=0}^{\infty} c_n x^n = c_0 + c_1 x + c_2 x^2 + ... + c_n x^n + ...$

Toute série entière est convergente en son centre; il existe des séries entières qui convergent dans tout le plan complexe $\left(\text{ainsi la série exponentielle } \sum_{n=0}^{\infty} \frac{x^n}{n!}\right)$, d'autres qui convergent en certains points et seulement en ceux-là (par exemple, la série géométrique convergente à l'intérieur du disque unité $|x| < 1$), d'autres enfin qui ne convergent qu'au centre $\left(\text{par exemple, la série } \sum_{n=0}^{\infty} n!\, x^n\right)$.

On montre que, lorsqu'une série entière converge en d'autres points que son centre — mais sans converger dans le plan complexe tout entier —, alors il existe un cercle, appelé cercle de convergence, centré au point *a*, tel que la série converge en tout point *x intérieur* à ce cercle et diverge en tout point *extérieur* à ce cercle (*fig.* 28). Le rayon de ce cercle est dit *rayon de convergence* de la série. Sur les points de la circonférence, le comportement de la série ne peut être établi de façon générale : il varie selon les cas. Une série entière qui converge dans tout le plan est dite avoir un *rayon de convergence infini*, tandis qu'une série entière ne convergeant qu'en son centre a un *rayon de convergence nul*.

La propriété essentielle du cercle de convergence d'une série entière est que cette série *converge uniformément* dans tout disque fermé de même centre et de rayon strictement plus petit que le rayon de convergence. Il y a même convergence absolue de la série en tout point situé à l'intérieur du cercle de convergence.

Le rayon de convergence dépend de la suite des coefficients c_n, en fait de la suite de leurs modules $|c_n|$. Son calcul s'effectue selon la classique *relation de Hadamard :*

$$\frac{1}{R} = \limsup_{n \to \infty} \sqrt[n]{|c_n|}$$

dans laquelle sont inclus, avec une interprétation évidente, le cas où $R = 0$ et le cas où $R = \infty$.

A l'intérieur du cercle de convergence, on peut aussi bien dériver terme à terme qu'intégrer terme à terme une série entière. La série des dérivées et la série des primitives ont d'ailleurs le même rayon de convergence.

L'étude des séries entières de rayon de convergence non nul a permis à K. Weierstrass de développer l'étude de la théorie des fonctions analytiques de variable complexe.

D'autres types de séries de fonctions ont une utilisation courante, tant en analyse mathématique que dans ses applications : les séries trigonométriques, dont on parlera plus loin pour les développements de fonctions en série de Fourier; les développements en série de fonctions orthogonales; les séries de Dirichlet, particulièrement intéressantes dans les récents développements de la théorie analytique des nombres; ou encore les séries de Laurent, à la base de la théorie des fonctions de variable complexe, pour ne citer que les plus classiques.

Équations différentielles

Il arrive souvent qu'à l'occasion de problèmes (géométriques, physiques, etc.) dans lesquels interviennent une variable indépendante *x* et une variable *y* dépendant de *x*, on ait à considérer — en vue de la solution — une équation liant *x*, *y* et ses dérivées (par rapport à *x*) $y', y'', ..., y^{(n)}$ jusqu'à un certain ordre *n* :

$$\varphi\,(x, y, y', ..., y^{(n)}) = 0.$$

De l'étude de cette relation, on se propose de déduire le maximum possible d'informations sur la dépendance de *y* selon *x*, plus précisément de déterminer la fonction $y = y\,(x)$ qui vérifie, quelle que soit la valeur prise par *x* dans un certain intervalle, l'équation :

$$\varphi\,(x, y\,(x), y'\,(x), ..., y^{(n)}\,(x)) = 0.$$

L'équation $\varphi\,(x, y, y', ..., y^{(n)}) = 0$ s'appelle *équation différentielle d'ordre n; n* est ici le plus grand ordre de dérivation parmi les dérivées figurant effectivement dans l'équation (par exemple, $x^2y''' - e^x \cdot y' + (\log x)\, y = 0$ est une équation différentielle d'ordre 3).

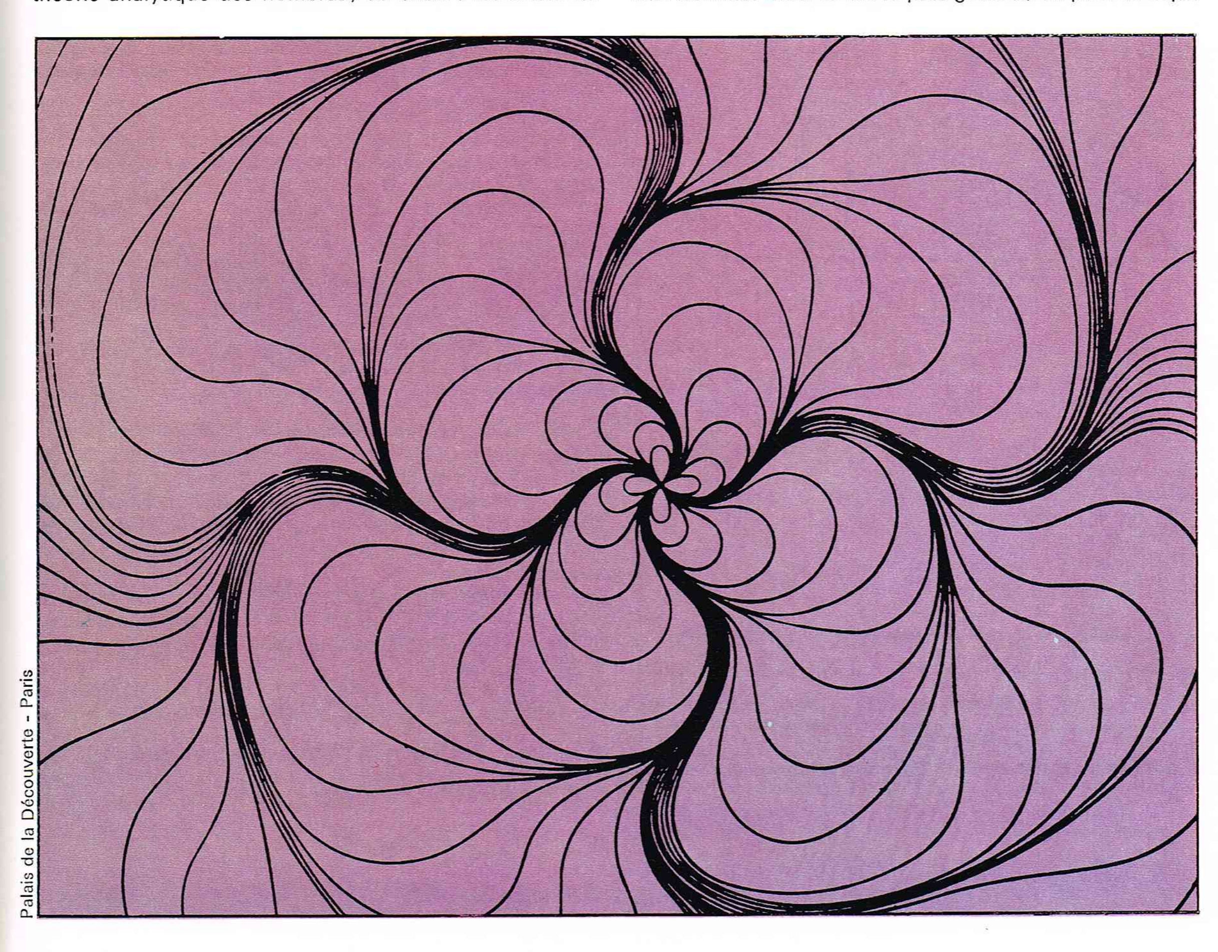
Lorsqu'il est possible de transformer l'équation de telle sorte que l'on puisse *résoudre* par rapport à (c'est-à-dire isoler) la dérivée de plus grand ordre, l'équation obtenue : $y^{(n)} = F\,(x, y, y', ..., y^{(n-1)})$ est dite écrite sous sa « forme normale »; il est évident que dans certains cas, une équation différentielle donne lieu à plusieurs « formes normales » :
par exemple, l'équation

$$y'''^2 - xy' + e^x = 0$$

peut s'écrire sous les deux formes normales

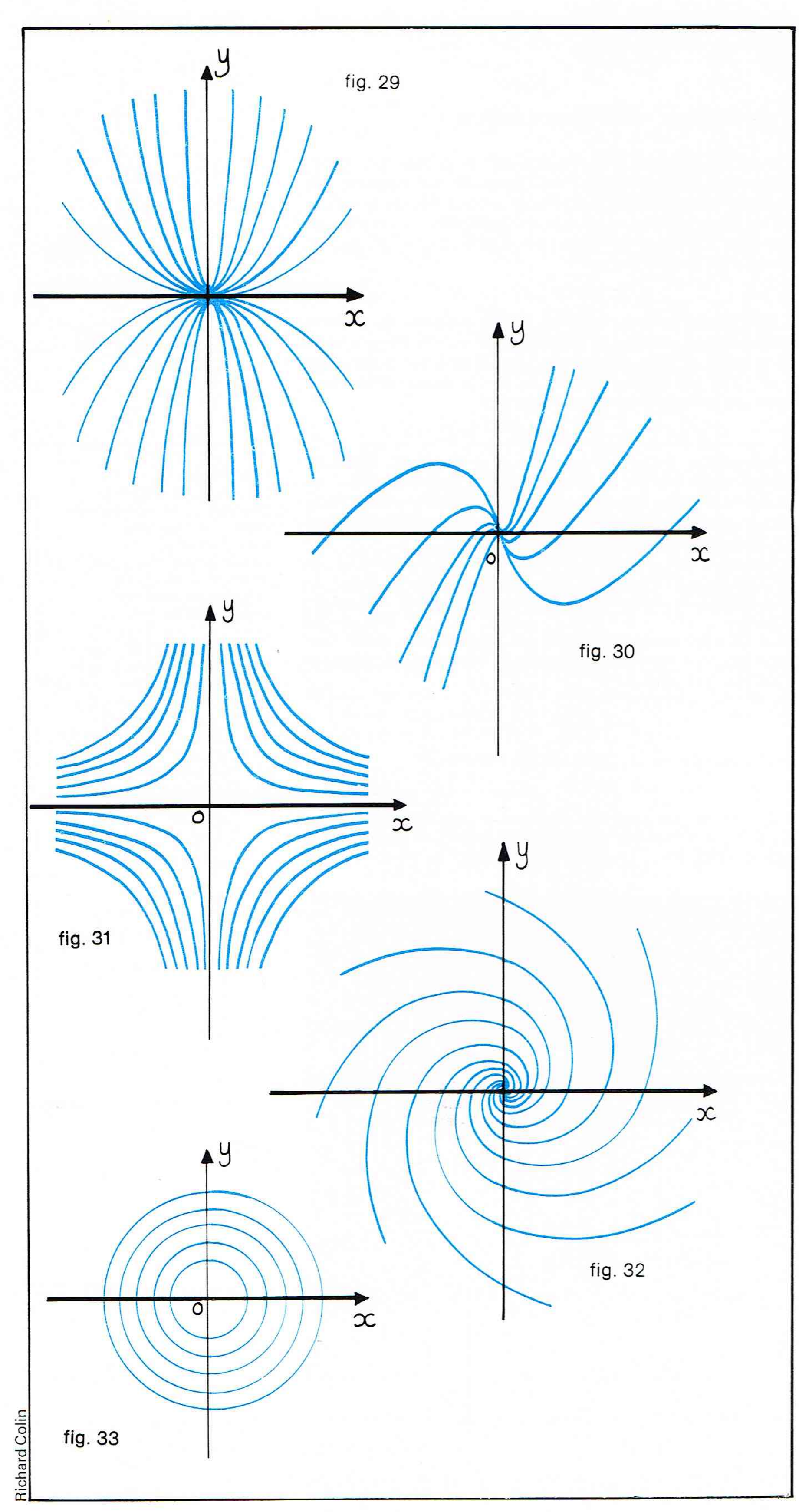
$$y''' = (xy' - e^x)^{1/2} \quad \text{ou} \quad y''' = -\,(xy' - e^x)^{1/2}.$$

Résoudre une équation différentielle signifie rechercher l'ensemble de toutes les fonctions *y* (*x*) vérifiant l'équation donnée. Dans le cas le plus général, on peut dire que



◀ ***Famille de solutions d'équation différentielle du premier degré.***

Palais de la Découverte - Paris



▲ *Figure 29 : famille des solutions* ($y = Cx^2$) *de l'équation* $xy' = 2y$.
Figure 30 : famille des solutions ($y = x \operatorname{Log}|x| + Cx$) *de l'équation* $xy' = x + y$.
Figure 31 : famille des solutions ($xy = C$) *de l'équation* $xy' = -y$ *avec pour* $C = 0$, $\{x = 0, y = 0\}$.
Figure 32 : famille des solutions ($r = Ce^{\varphi}$) *de l'équation* $y' = \dfrac{x+y}{x-y}$ *avec* $x = r \cos \varphi$, $y = r \sin \varphi$.
Figure 33 : famille des solutions ($x^2 + y^2 = C$) *de l'équation* $yy' = -x$.

lorsqu'une équation différentielle d'ordre n possède des solutions, celles-ci sont définies implicitement par une famille d'équations du type $\Psi(x, y, c_1, c_2, ..., c_n) = 0$, et le **théorème des fonctions implicites** permet de conclure à l'existence d'un ou plusieurs systèmes de fonctions $y = y(x; c_1, c_2, c_3, ..., c_n)$, chacune dépendant en fait de n paramètres arbitraires.

En adoptant un langage inspiré de considérations à caractère géométrique, on appellera *courbe intégrale de l'équation* une telle famille de solutions; en effet, on pourrait construire (par dérivation et élimination de paramètres) une équation différentielle d'ordre m, à partir d'une famille de courbes $\varphi(x, y; a_1, a_2, ..., a_m) = 0$ dépendant effectivement de m paramètres. Par exemple, en partant de la famille des cercles de rayon r fixe et de centre (a, b) variable dans le plan :

$$(x - a)^2 + (y - b)^2 - r^2 = 0$$

on obtient en dérivant deux fois (autant de fois que de paramètres) :

$$(x - a) + (y - b)\, y' = 0$$

puis $$1 + y'^2 + (y - b)\, y'' = 0$$

et l'élimination des paramètres a, b entre ces trois équations donne l'équation différentielle du second ordre :

$$(1 + y'^2)^3 - r^2 y''^2 = 0$$

qui admet comme système de courbes intégrales la famille (doublement infinie) des cercles $\varphi(x, y; a, b) = 0$ de laquelle nous sommes partis, et qui dépend de deux paramètres arbitraires.

L'équation $\Psi(x, y; c_1, c_2, ..., c_n) = 0$ s'appelle *l'intégrale générale* de l'équation, et pour un *n-uplet* particulier de constantes : $c_1, c_2, ..., c_n$, on obtient $\Psi(x, y; c_1, c_2, ..., c_n)$ que l'on appelle une *intégrale particulière ;* les *intégrales singulières* sont, elles, des fonctions $y = \gamma(x)$ vérifiant l'équation différentielle mais ne pouvant être considérées comme des cas particuliers de l'intégrale générale.

L'expression exacte de l'intégrale générale par des opérations et fonctions usuelles est réalisable, à un niveau élémentaire, uniquement pour quelques types particuliers d'équations, présentées plus loin. Naturellement, dès que l'on dispose de notions et d'instruments d'analyse mathématique plus « sophistiqués », la classe des équations différentielles résolubles (au sens indiqué) s'élargit beaucoup.

Dans les autres cas, et selon les contraintes du problème pratique de départ, avant de rechercher l'ensemble de toutes les solutions, on essaie de trouver une solution particulière possédant les propriétés demandées par le problème particulier qui est examiné : ces propriétés peuvent avoir trait au comportement de la solution particulière en un point, par exemple aux « conditions initiales »; la détermination d'une solution vérifiant un système donné de conditions initiales est désignée sous le nom de « *problème de Cauchy* ». D'autres fois, les conditions restrictives peuvent être relatives à des points divers, et l'on dit alors que l'on a un « *problème aux limites* ».

Types simples d'équations du premier ordre

Parmi les équations du premier ordre (*fig.* 29, 30, 31, 32 et 33), on distinguera les suivantes :

● *Les équations à variables séparées ou séparables*, que l'on peut donc mettre sous la forme $A(x)\, dx = B(y)\, dy$, qui donnent alors $\int A(x)\, dx = \int B(y)\, dy + c$ pour intégrale générale.

● *Les équations différentielles exactes*, de la forme $A(x, y)\, dx + B(x, y)\, dy = 0$ où le premier membre est la différentielle exacte d'une certaine « fonction potentiel » $U(x, y)$; il faut donc que

$$\frac{\partial A}{\partial y}(x, y) = \frac{\partial B}{\partial x}(x, y)$$

et $U(x, y) = c$ est alors l'intégrale générale.

● *Les équations homogènes*, qui s'écrivent $y' = f\left(\dfrac{y}{x}\right)$, où $y' = \dfrac{dy}{dx}$; la transformation $y = t \cdot x$

(donc $dy = t \cdot dx + x dt$)

ramène alors ce type d'équation au type précédent.

• *Les équations linéaires,* de la forme

$$y' = A(x)\, y + B(x).$$

L'intégrale générale de cette équation s'obtient en ajoutant à une solution particulière l'intégrale générale de l'équation homogène $y' = A(x)\, y$ associée, et s'écrit sous la forme :

$$y = \exp\left(\int A(x)\, dx\right) \cdot \left[B(x) \exp\left(-\int A(x)\, dx + c\right)\right];$$

la méthode utilisée est celle de la « variation de la constante ».

• *Les équations de Bernouilli :* $y' = A(x)\, y + B(x)\, y^r$, où r est un nombre réel non nul et différent de 1.

La substitution $y = z^{1/1-r}$ (donc $y' = \frac{1}{1-r} z^{\frac{r}{1-r}} \cdot z'$) transforme l'équation donnée en une équation différentielle linéaire, du type précédent, de la variable $z = z(x)$;

• *Les équations de Riccati :*

$$y' = A(x)\, y^2 + B(x)\, y + C(x).$$

La résolution présuppose la connaissance d'une solution particulière $\varphi(x)$; la transformation $y = \varphi(x) + \frac{1}{z}$ ramène alors l'équation donnée à une équation différentielle linéaire relativement à $z = z(x)$.

Dans les quatre derniers types d'équations, il s'agissait d'équations résolues par rapport à y' ; les deux suivantes sont, par contre, des cas que l'on ne peut mettre sous forme normale.

• *L'équation de Lagrange :*

$$y = xA(y') + B(y') \qquad [B(y') \neq y'].$$

En posant $y' = t = t(x)$, on obtient $y = xA(t) + B(t)$; en dérivant, il vient $t = A(t) + xA'(t) \cdot t' + B'(t) \cdot t'$ avec $t' = \frac{dt}{dx}$; d'où l'équation linéaire de la variable $x = x(t)$:

$$\frac{dx}{dt} = \frac{A'(t)}{t - A(t)} \cdot x + \frac{B'(t)}{t - A(t)}$$

• *L'équation de Clairaut :* $y = xy' + A(y')$. On pose $y' = t$ et on obtient l'intégrale générale $y = cx + A(c)$. Intégrale singulière : $\begin{cases} x = -A'(t) \\ y = tx + A(t) \end{cases}$ que l'on obtient simplement en dérivant l'équation de départ par rapport à x.

Types simples d'équations du second ordre

• Dans les équations de la forme $F(x, y', y'') = 0$, on pose $y' = t = t(x)$ donc $y'' = t'$. Ceci ramène l'équation à une équation du premier ordre relativement à t de laquelle résulte l'intégrale $\varphi(x, t, c_1) = 0$; par conséquent, on a de nouveau une équation du premier ordre $\varphi(x, y', c_1) = 0$ de laquelle on déduit l'intégrale générale $f(x, y; c_1, c_2) = 0$.

• Dans les équations de la forme $F(y, y', y'') = 0$ on tient compte de l'égalité $y'' = \frac{dy'}{dy} \cdot y'$ pour considérer la transformation $y' = t(y)$. On obtient alors une équation du premier ordre : $F(y, t, t' \cdot t) = 0$ dont l'intégrale générale a la forme $\varphi(y, t; c_1) = 0$. Puis de cette nouvelle équation différentielle du premier ordre $\varphi(y, y'; c_1) = 0$, on obtient l'intégrale générale cherchée sous la forme : $f(x, y; c_1, c_2) = 0$ (*fig.* 34, 35, 36 et 37).

Équations différentielles linéaires d'ordre n

Il s'agit d'équations de la forme :

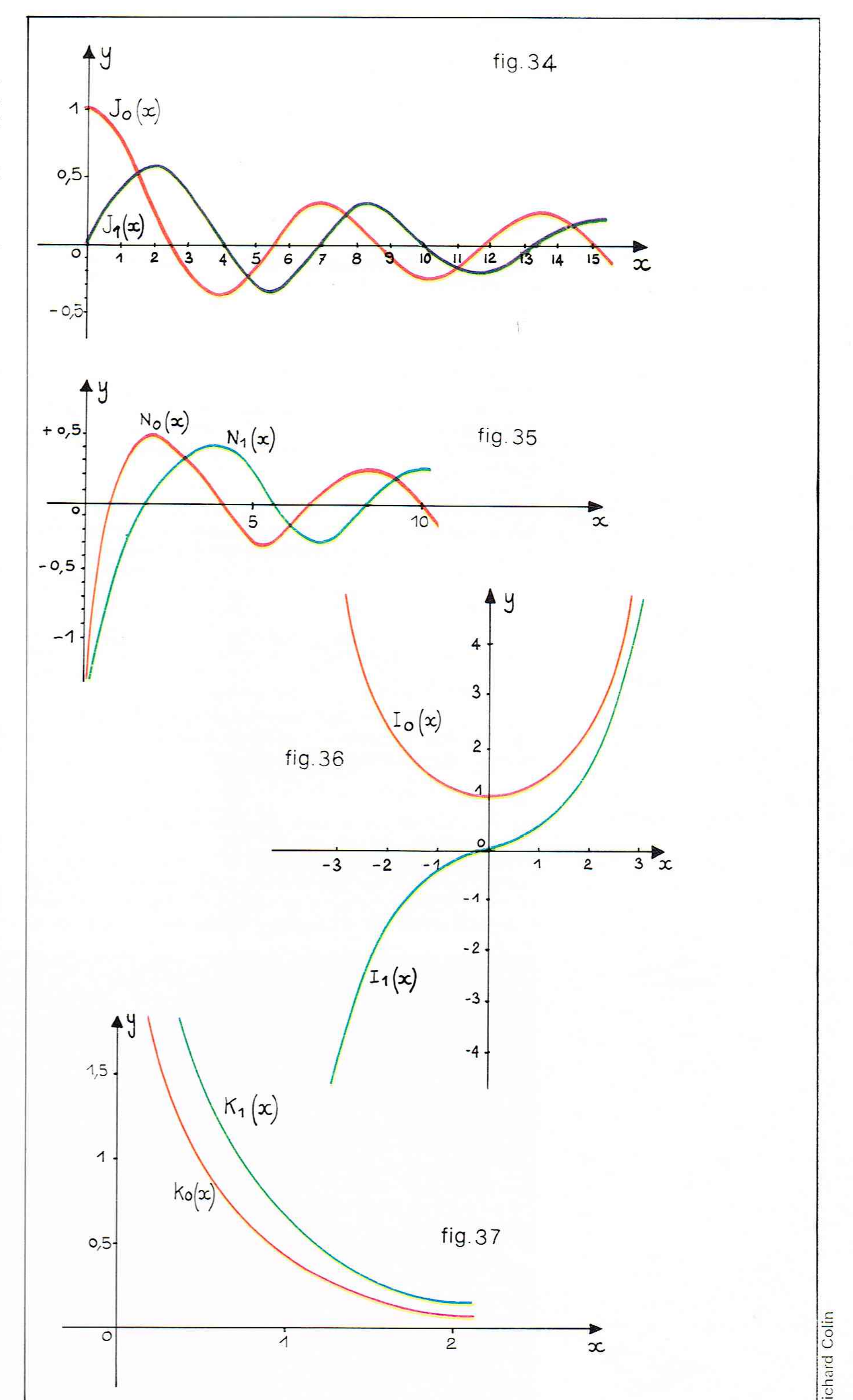
$$p_0(x)\, y^{(n)} + p_1(x)\, y^{(n-1)} + \ldots + p_n(x)\, y = A(x)$$

où les fonctions $p_i(x)$, $(i = 1, 2, \ldots, n)$ et $A(x)$ sont continues dans un intervalle $[a, b]$, avec de plus $p_0(x) \neq 0$ pour tout $x \in [a, b]$. Lorsque le second membre $A(x)$ n'est pas identiquement nul sur $[a, b]$, on dit que l'équation est complète, et si $A(x) = 0$ sur $[a, b]$, on dit que l'on a une équation homogène.

La linéarité du premier membre, relativement à

$$y, y', \ldots, y^{(n)}$$

entraîne que, si $\varphi_1(x), \varphi_2(x), \ldots, \varphi_k(x)$ représentent des solutions quelconques de l'équation homogène (c'est-à-



▲ *Figure 34 : les deux premières* **fonctions de Bessel** *$J_0(x)$ et $J_1(x)$, solutions de l'équation :* $y'' + \frac{1}{x} y' + \left(1 - \frac{\nu^2}{x^2}\right) y = 0$ *pour* $\nu = 0$ *et* $\nu = 1$.

Figure 35 : les deux premières **fonctions de von Neumann,** *définies par*
$N_\nu(x) = \frac{\cos \nu\pi \cdot J_\nu(x) - J_{-\nu}(x)}{\sin \nu\pi}$, $\nu \neq \mathbb{Z}$
$N_k(x) = \lim_{\nu \to k} N_\nu(x)$, $k \in \mathbb{Z}$ *pour* $k = 0$ *et* $k = 1$.

Figure 36 : les fonctions $I_0(x)$ et $I_1(x)$, solutions de $x^2y'' + xy' + (x^2 + n^2)\, y = 0$ *pour* $n = 0$ *et* $n = 1$*, sont liées aux fonctions de Bessel $J_0(x)$ et $J_1(x)$ par* $I_n(x) = i^{-n} \cdot J_n(ix)$.

Figure 37 : les deux premières **fonctions de Kelvin** *$K_0(x)$ et $K_1(x)$ définies par* $K_n(x) = \frac{\pi}{2} \cdot \frac{I_{-n}(x) - I_n(x)}{\sin n\pi}$ *pour* $n = 0$ *et* $n = 1$.

dire sans second membre), alors toute combinaison linéaire de ces solutions,

$$y(x) = c_1 \varphi_1(x) + c_2 \varphi_2(x) + \dots + c_k \varphi_k(x)$$

(où les c_i sont des constantes) est encore solution de la même équation homogène.

On montre que l'ensemble des solutions d'une équation différentielle linéaire d'ordre n forme, pour les opérations usuelles d'addition et de produit par une constante, un espace vectoriel de dimension finie n. Il est donc suffisant, pour construire l'intégrale générale d'une telle équation, de déterminer exactement n solutions linéairement indépendantes, puisque celles-ci forment alors une base de l'espace des solutions : on dit que l'on a un système fondamental de solutions : $\varphi_1(x)$, $\varphi_2(x)$, ..., $\varphi_n(x)$; la famille des fonctions $c_1\varphi_1(x) + c_2\varphi_2(x) + \dots + c_n\varphi_n(x)$ dépendant des n paramètres $c_1, c_2, \dots, c_n$ est alors l'intégrale générale de l'équation homogène donnée.

Dans le cas des équations différentielles linéaires complètes, on montre que l'intégrale générale s'obtient en ajoutant à une intégrale particulière $y^*(x)$ la solution générale de l'équation linéaire homogène associée (c'est-à-dire où l'on a remplacé le second membre A (x) par 0) ; donc :

$$y = y^*(x) + c_1\varphi_1(x) + c_2\varphi_2(x) + \dots + c_n\varphi_n(x).$$

Un cas particulier très intéressant et fréquemment rencontré est celui des équations différentielles linéaires à coefficients constants :

$$a_0 y^{(n)} + a_1 y^{(n-1)} + \dots + a_n y = Q(x)$$

$a_0, a_1, \dots, a_n$ étant des constantes. On considère l'équation homogène associée :

$$a_0 y^{(n)} + a_1 y^{(n-1)} + \dots + a_n y = 0.$$

La recherche d'un système fondamental de solutions, dans ce cas, conduit à la résolution de l'équation algébrique, dite *équation caractéristique :*

$$a_0 r^n + a_1 r^{n-1} + \dots + a_n = 0$$

où les coefficients a_i sont les mêmes que dans l'équation proposée, et r l'inconnue.

Dans le domaine complexe $\mathbb{C}$, une telle équation admet **(théorème de D'Alembert-Gauss)** n racines $r_1, r_2, \dots, r_n$; les fonctions exp $(r_k x)$ $(k = 1, 2, \dots, n)$ sont des intégrales de l'équation différentielle homogène, puisque :

$$[\exp(r_k x)]^{(p)} = r_k^p \exp(r_k x).$$

Si les racines r_k sont toutes distinctes, les n intégrales particulières $e^{r_k x}$ $(k = 1, 2, \dots, n)$ constituent un système fondamental de solutions, et, par suite, l'intégrale générale qui en résulte est :

$$y = c_1 e^{r_1 x} + c_2 e^{r_2 x} + \dots + c_n e^{r_n x}$$

où $c_1, c_2, \dots, c_n$ sont des *constantes arbitraires*.

Si l'une des racines r est multiple et si m $(\leqslant n)$ désigne son ordre de multiplicité, alors la suite (le *m-uple*) de solutions :

$$e^{rx}, xe^{rx}, x^2e^{rx}, \dots, x^m e^{rx}$$

forme un système linéairement indépendant, qui, réuni avec les $n - m$ racines distinctes restantes, constitue un système fondamental de solutions, et l'intégrale générale cherchée est combinaison linéaire des éléments de ce système.

Lorsque l'on est dans le cas où l'équation caractéristique admet des racines complexes, bien que les coefficients a_i soient réels, on remarque que pour toute racine complexe $r = \alpha + i\beta$, le nombre complexe conjugué $\bar{r} = \alpha - i\beta$ est aussi racine, de telle sorte que, par combinaison linéaire appropriée des deux solutions $e^{(\alpha + i\beta)x}$ et $e^{(\alpha - i\beta)x}$, et grâce aux formules d'Euler (voir *Trigonométrie*), on obtienne les deux solutions particulières réelles :

$$e^{\alpha x} \cos \beta x \quad \text{et} \quad e^{\alpha x} \sin \beta x.$$

Équations aux dérivées partielles

Lorsque la fonction inconnue est une fonction de deux ou plusieurs variables indépendantes, l'équation qui lie la fonction à ses dérivées, donc en fait les dérivées partielles, jusqu'à l'ordre n, s'appelle *équation aux dérivées partielles* d'ordre n. Ces équations sont fondamentales, par exemple en physique mathématique, en physique théorique (où les problèmes traités admettent usuellement le temps et les coordonnées d'un point arbitraire comme variables indépendantes), en économie mathématique, etc.

La résolution de ces équations est, en général, délicate et complexe. Après avoir posé le problème et quelques considérations de base, nous examinerons ici quelques équations classiques.

● Les *équations aux dérivées partielles du premier ordre*, pour des fonctions de deux variables indépendantes, se présentent sous la forme :

$$\varphi\left(x, y, z, \frac{\partial z}{\partial x}, \frac{\partial z}{\partial y}\right) = 0.$$

Or, il est à constater que, lors d'une dérivation partielle, comme pour une constante additive, on a « disparition » de tout ce qui est fonction de l'autre variable (par exemple, si $z = y^2 + \Psi(x)$, on a $\frac{\partial z}{\partial y} = 2y$) ; on voit là le moyen d'obtenir une équation aux dérivées partielles par différence d'équations différentielles classiques au moyen d'une élimination appropriée de fonctions arbitraires. Les solutions d'équations aux dérivées partielles contiendront donc des *fonctions arbitraires*, de même que les solutions d'équations différentielles contiennent des constantes arbitraires.

Ainsi, l'équation aux dérivées partielles du premier ordre, linéaire, s'écrit :

$$A(x, y, z)\frac{\partial z}{\partial x} + B(x, y, z)\frac{\partial z}{\partial y} = C(x, y, z)$$

où A, B, C, sont des fonctions données [une telle équation pourrait être obtenue de l'équation $\varphi(u, v) = 0$, où $u = u(x, y, z)$ et $v = v(x, y, z)$, en éliminant la fonction arbitraire φ].

● Parmi les *équations aux dérivées partielles d'ordre supérieur à* 1, fréquemment utilisées dans tous les domaines de la physique, on peut distinguer :

— *L'équation des cordes vibrantes :*

$$\frac{\partial^2 y}{\partial t^2} = a^2 \frac{\partial^2 y}{\partial x^2}$$

où t représente le temps ; la corde élastique homogène est tendue entre les points (0, 0) et (l, 0) de l'axe x, vibrant ensuite dans le plan Oxy ; l'abscisse d'un point de la corde x détermine au temps t son ordonnée $y = y(x, t)$; enfin a^2 est une constante liée aux caractéristiques physiques de la corde.

Le changement de variable $x + at = u$, $x - at = v$ transforme l'équation de départ en $\frac{\partial^2 y}{\partial u \, \partial v} = 0$, que l'on résout :

$$\frac{\partial y}{\partial u} = f_1(u), \text{ où } f_1 \text{ est arbitraire,}$$

soit $y = f(u) + g(v)$, où f est une primitive de f_1 et g une fonction arbitraire. D'où la solution générale de l'équation :

$$y(x, t) = f(x + at) + g(x - at)$$

où f et g sont deux fonctions (continûment différentiables) deux fois arbitraires. Il est aisé de donner un aspect physique à cette intégrale générale si l'on remarque que la constante a peut s'interpréter comme une vitesse de propagation des ondes.

Le problème de Cauchy, qui correspond aux conditions initiales

$$\begin{cases} y_0(x) = y(x, 0), \text{ forme de la corde à l'instant initial;} \\ y_1(x) = \frac{\partial y}{\partial t}(x, 0), \text{ vitesse le long de la corde à l'instant initial;} \end{cases}$$

admet une solution unique :

$$y(x, t) = \frac{1}{2}\left\{ y_0(x + at) + y_0(x - at) \right\} + \frac{1}{2a}\int_{x-at}^{x+at} y_1(u)\, du$$

La résolution de cette équation permet de discuter de nombreux problèmes physiques tels que ceux posés par des vibrations longitudinales ou par des tuyaux sonores.

— *Équation de Laplace :*

$$\Delta V = \frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} + \frac{\partial^2 V}{\partial z^2} = 0.$$

La fonction V (x, y, z) — où (x, y, z) désigne un point quelconque de l'espace — admet comme interprétations physiques les suivantes :

potentiel de gravitation en milieu privé de points matériels ;

potentiel électrique dans un diélectrique uniforme ;

potentiel magnétique, etc.

— *Équation de Poisson :*

$$\Delta V = \frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} + \frac{\partial^2 V}{\partial z^2} = f(x, y, z)$$

où $f(x, y, z)$ est une fonction donnée, définie par le problème concret.

Ici V peut être interprété comme :

potentiel de gravitation dans un milieu où $f(x, y, z)$ est proportionnel à la densité de matière ;

potentiel électrostatique dans un milieu où $f(x, y, z)$ est proportionnel à la distribution des charges, etc.

— *Équation de la chaleur :*

$$\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} + \frac{\partial^2 V}{\partial z^2} = \frac{1}{k} \frac{\partial V}{\partial t}$$

où V représente la température à l'instant t du point (x, y, z) d'un corps homogène isotrope, et k une constante de diffusion, caractéristique du matériau. Lorsque la température est stationnaire, alors $\frac{\partial V}{\partial t} = 0$, et l'on retrouve l'équation de Laplace.

Dans le cas de l'équation, dite simplifiée : $\frac{\partial^2 V}{\partial x^2} = \frac{\partial V}{\partial t}$, on montre que, pour toute fonction $f(u)$ suffisamment régulière, on a la solution :

$$G_f(x, t) = \int_{-\infty}^{+\infty} \frac{1}{2\sqrt{\pi t}} \exp\left(\frac{-(x-u)^2}{4t}\right) \cdot f(u)\, du$$

Le problème de Cauchy, correspondant à la condition initiale $V_0(x) = V(x, 0)$, admet la solution :

$$V(x, t) = \int_{-\infty}^{+\infty} \frac{1}{2\sqrt{\pi t}} \exp\left(\frac{-(x-u)^2}{4t}\right) \cdot V_0(u)\, du$$

— *Équation des ondes :*

$$\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} + \frac{\partial^2 V}{\partial z^2} = \frac{1}{c^2} \cdot \frac{\partial^2 V}{\partial t^2}$$

où c est la vitesse de propagation de l'onde.

Ses applications sont bien connues en théories électromagnétiques. Dans le cas monodimensionnel, on retrouve l'équation des cordes vibrantes.

— *Équation de Schrödinger :*

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = \frac{8\pi^2 m}{h^2}(V - W)\, u$$

où m désigne la masse d'une particule dotée de l'énergie totale W, V étant son énergie potentielle, et h la constante de Planck. Cette équation intervient en mécanique quantique.

Fonctions de variable complexe

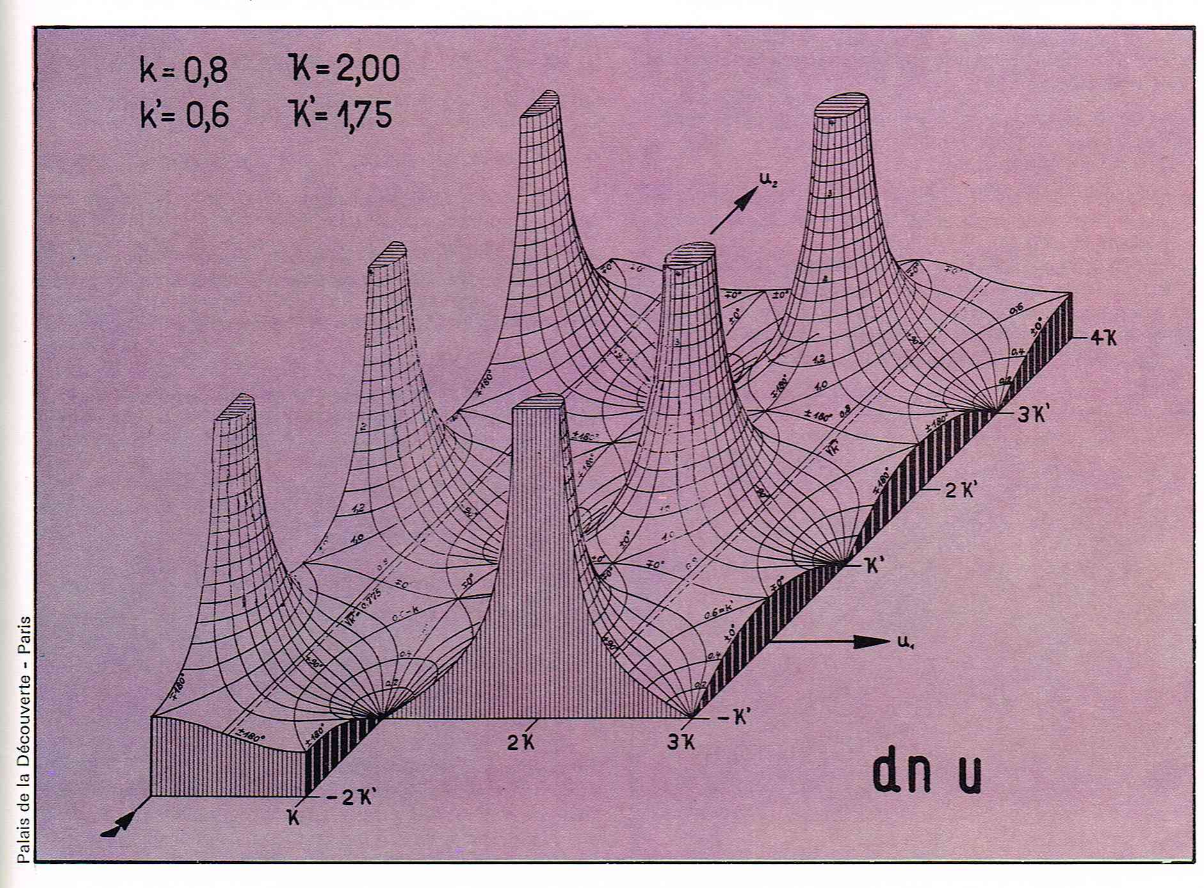
L'étude des séries entières est, historiquement, à l'origine du développement de l'analyse par les fonctions de variable complexe ; c'est d'abord une extension du cadre des fonctions analytiques (d'où résultent le concept essentiel de dérivation complexe et celui de fonction holomorphe), puis l'étude plus générale menée à l'aide de l'intégrale curviligne : on aborde alors tous les problèmes de primitives. La recherche de solutions globales, étendant les résultats locaux obtenus par la théorie de Cauchy, débouche alors rapidement sur des problèmes de géométrie et de topologie algébrique.

Sur un plan très pratique, l'un des aboutissements de la théorie de Cauchy est le **théorème des résidus** qui est continuellement utilisé en calcul intégral ; il n'est pas exagéré de dire qu'il s'agit d'un des outils mathématiques indispensables à la physique classique.

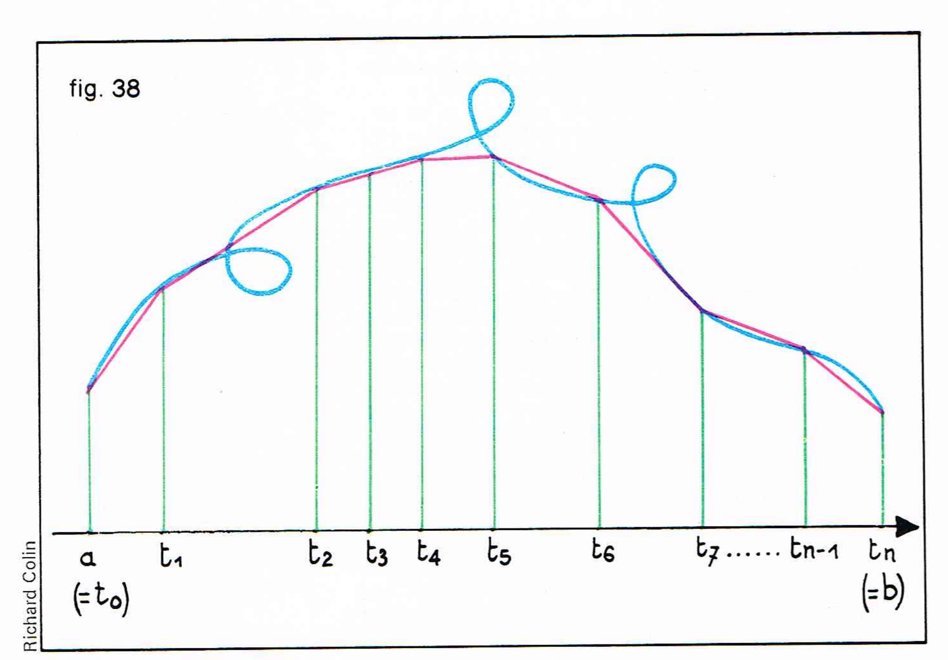
Loin d'être épuisé, le sujet fait encore l'objet de très nombreuses recherches en liaison étroite avec les résultats acquis en topologie algébrique.

Fonctions holomorphes

La dérivation des fonctions de variable complexe est, comme dans le cas réel, l'outil de base de l'analyse. Mais la différence est grande, et le concept de dérivée par rapport à la variable complexe $z = x + iy$ est particulièrement puissant, puisque en fait il recouvre beaucoup plus



◄ ***Relief de la fonction elliptique :*** **dn (u) = $V_1 - K_2 SN_2$ (u).**



▲ *Figure 38 : exemple de courbe rectifiable.*

complètement que dans le cas réel l'approximation des valeurs prises par une fonction au voisinage d'un point. La *formule de Taylor,*

$$f(a+h) = f(a) + h f'(a) + \frac{h^2}{2!} f''(a) + \ldots + \frac{h^n}{n!} f^{(n)}(a) + \frac{h^{n+1}}{(n+1)!} R_n$$

exprimée pour les fonctions réelles d'une variable réelle, permet cette approximation qui est d'autant plus précise que la fonction est à un ordre élevé continûment dérivable.

Mais, malgré ces conditions restrictives, il est aisé de trouver des fonctions « assez régulières » qui ne coïncident pas avec leur développement de Taylor :

$$\sum^{\infty} \frac{1}{n!} f^{(n)}(x_0) \cdot (x - x_0)^n$$

autour d'un point x_0 donné ; par exemple, la fonction définie sur $\mathbb{R}$ par :

$$\begin{cases} f(x) = \exp\left(-\frac{1}{x^2}\right) & \text{si } x \neq 0 \\ f(0) = 0 \end{cases}$$

dont la série de Taylor au point $x_0 = 0$ a tous ses coefficients nuls et converge donc, mais n'est égale à f qu'au seul point $x_0 = 0$.

On dit qu'une *fonction* est *analytique* sur un ouvert $U \subset R$, si elle est égale à son développement de Taylor au point x_0, au voisinage de tout point $x_0 \in U$. Il en est ainsi de multiples fonctions usuelles : logarithme, exponentielle, fonctions trigonométriques circulaires ou hyperboliques, etc. La classe de ces fonctions va se révéler plus aisée à déterminer dès que l'on s'intéressera aux variables complexes.

Si donc $f : \Omega \to \mathbb{C}$, où Ω est un ouvert de $\mathbb{C}$, est une fonction de variable complexe, on dit que f est *dérivable* en $a \in \Omega$ si l'expression

$$\lim_{\substack{h \to 0 \\ \neq 0 \\ h \in \mathbb{C}}} \frac{f(a+h) - f(a)}{h} \quad \text{existe,}$$

(on la note $f'(a)$) en posant $h = h_1 + ih_2$, tel que h tend vers zéro si et seulement si h_1 et h_2 (réels) tendent vers zéro. On dira que f est dérivable ou encore *holomorphe* sur Ω, si ceci est vrai en tout point $a \in \Omega$. Les conditions de Cauchy-Riemann sont des conditions nécessaires et suffisantes d'holomorphie. Si l'on décompose $f(z) = f(x + iy) = P(x, y) + iQ(x, y)$, comme tout nombre complexe, en partie réelle et partie imaginaire, elles s'écrivent :

$$\frac{\partial P}{\partial x} = \frac{\partial Q}{\partial y}; \qquad \frac{\partial P}{\partial y} = -\frac{\partial Q}{\partial x}$$

Bien entendu, comme pour les fonctions réelles, dérivables, d'une variable réelle, la somme et le produit de fonctions holomorphes, l'inverse d'une fonction holomorphe non nulle sont holomorphes ; le produit de composition — lorsqu'il est licite ($f : \Omega \to G$ et $g : G \to \mathbb{C}$ ont un produit de composition $g \circ f : \Omega \to \mathbb{C}$) — de deux fonctions holomorphes est holomorphe. Dans tous les cas, les formules usuelles de dérivation se transcrivent sans modification.

Si l'on définit, pour $z = x + iy$ et une fonction $f(z)$, les dérivées :

$$\frac{\partial f}{\partial z} = \frac{1}{2}\left(\frac{\partial f}{\partial x} - i\frac{\partial f}{\partial y}\right) \quad \text{et} \quad \frac{\partial f}{\partial \bar{z}} = \frac{1}{2}\left(\frac{\partial f}{\partial x} + i\frac{\partial f}{\partial y}\right)$$

qui se justifient par le fait que :

$$\begin{cases} dz = dx + idy \\ d\bar{z} = dx - idy \end{cases}$$

donc que :

$$dx = \frac{1}{2}(dz + d\bar{z}) \quad \text{et} \quad dy = \frac{1}{2i}(dz - d\bar{z}),$$

donc aussi que :

$$df = \frac{\partial f}{\partial z} dz + \frac{\partial f}{\partial \bar{z}} d\bar{z} \quad \text{puisque } df = \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy,$$

alors la fonction f est holomorphe si et seulement si on a

$$\frac{\partial f}{\partial \bar{z}} = 0.$$

La fonction $z \to z$ est holomorphe, donc les polynômes et les fractions rationnelles (en dehors de leurs pôles, c'est-à-dire des points où s'annule le dénominateur) sont holomorphes. Pour une série entière $\sum^{\infty} a_n z^n$, ayant un rayon de convergence égal à R (qui peut aussi être infini), la fonction $z \to \sum^{\infty} a_n z^n$ qu'elle définit est *holomorphe à l'intérieur* du disque de convergence, c'est-à-dire pour tout $z \in \mathbb{C}$ tel que $|z| < R$.

Par suite, les fonctions

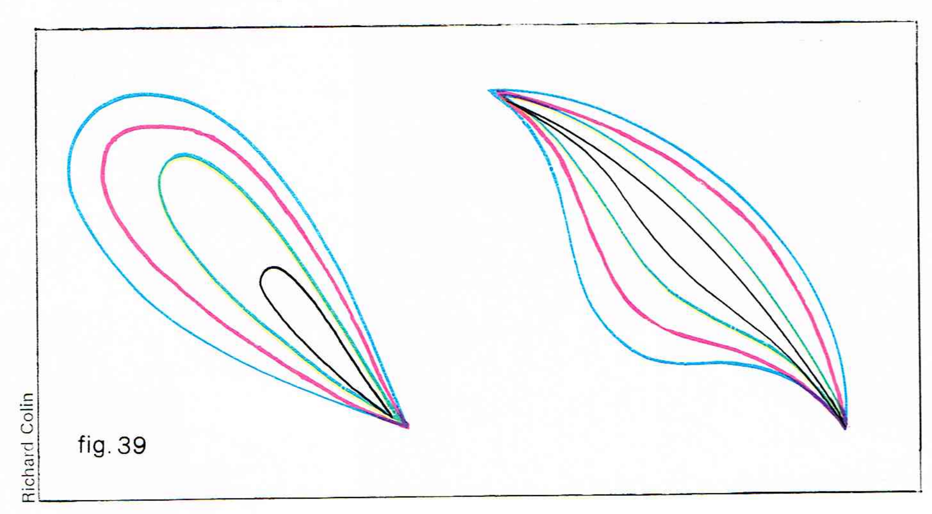
$$\begin{cases} z \to e^z = \sum_{n=0}^{\infty} \frac{z^n}{n!} \\ z \to \sin z = \sum_{n=0}^{\infty} (-1)^n \frac{z^{(2n+1)}}{(2n+1)!} \\ z \to \cos z = \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n}}{(2n)!} \end{cases}$$

sont holomorphes dans le plan complexe tout entier.

Par contre, les fonctions $z \to \bar{z}$, ou encore $z \to \arg z$, ne sont pas holomorphes.

La notion d'holomorphie, apparemment identique à celle de dérivabilité pour les fonctions réelles d'une variable réelle, est en fait beaucoup plus riche, et par conséquent est un concept mathématique très fort.

▼ *Figure 39 : notion d'homotopie.*



Intégrales curvilignes — Théorie de Cauchy

Pour tout segment $[a, b] \subset \mathbb{R}$ et toute application continue $\varphi : [a, b] \to \mathbb{C}$, on appelle *courbe* le couple formé par l'application φ et son image $\varphi([a, b])$, soit $\Gamma = \{\varphi([a, b]), \varphi\}$. On dit que cette courbe est continûment différentiable par morceaux si l'application φ possède cette propriété. On donne le nom de *paramétrage* de Γ à φ. Pour une même courbe Γ, on peut envisager des paramétrages distincts ; toutefois, deux paramétrages φ et Ψ de classe C^n seront dits *équivalents :*

$$\begin{cases} \varphi : [\alpha, \beta] \to \mathbb{C} \\ \Psi : [a, b] \to \mathbb{C} \end{cases}$$

pourvu qu'il existe une application $h : [\alpha, \beta] \to [a, b]$, bijective, strictement croissante, de classe C^n telle que : $\varphi = \Psi \circ h$, avec de plus $h'(t) \neq 0$ pour tout $t \in [\alpha, \beta]$.

Ajoutons enfin que, lorsque la fonction de paramétrage φ est telle que, pour toute division

$$a = t_0 < t_1 < t_2 < ... < t_n = b$$

de l'intervalle $[a, b]$, on a $\sum_{i=1}^{n} |\varphi(t_i) - \varphi(t_{i-1})| < C_\varphi$

où C_φ est une constante qui ne dépend que de φ et non pas du partage $|t_i|$ — on dit que φ est à variation bornée — la *courbe* Γ est dite *rectifiable* (*fig.* 38).

Dans ce cas, $\sum_{i=1}^{n} |\varphi(t_i) - \varphi(t_{i-1})| = L_{\{ti\}}(\varphi)$ représente la longueur de la ligne polygonale inscrite dans Γ, définie par les points M_i ($i = 1, 2, ..., n$) de la courbe, d'abscisse t_i ($i = 1, 2, ..., n$). Le pas du partage $\{t_i\}$ étant défini naturellement par $\delta(P) = \sup_i |t_i - t_{i-1}|$ (c'est donc $t_2 - t_1$ dans le cas de figure), si $L_{\{ti\}}(\varphi)$ admet une limite lorsque $\delta(P) \to 0$, cette limite est, bien sûr, la *longueur* de la courbe :

$$\text{longueur } \Gamma = \int_a^b |d\,\varphi(t)| = \int_a^b |\varphi'(t)|\, dt$$

$$= \sup_{\text{partage } \{t_i\}} \sum_{i=1}^{n} |\varphi(t_i) - \varphi(t_{i-1})|.$$

Le dernier terme est appelé *variation* de φ.

Si f est une application d'un ouvert $\Omega \subset \mathbb{C}$, dans $\mathbb{C}$, continue, et si Γ est une courbe paramétrée par φ et $[a, b]$, d'image γ, alors l'expression $\int_a^b f[\varphi(t)]\, d\varphi(t)$ est invariante dans tout changement de paramétrage à l'intérieur de la classe des paramètres équivalents ; on la désigne par la notation $\int_\gamma f(z)\, dz$. En particulier, la longueur de γ se note $\int_\gamma |dz|$.

C'est alors tout le problème des primitives, au sens complexe, que l'on peut poser à l'aide des *notions de lacet et d'homotopie* (*fig.* 39), et d'où on peut alors faire ressortir le caractère très particulier des fonctions holomorphes.

Un *lacet* est un chemin, c'est-à-dire une application $\gamma : I = [a, b] \to \mathbb{C}$, dont les extrémités sont égales : $\gamma(a) = \gamma(b)$. Lorsque l'application γ est constante, le *lacet* est dit *constant;* sa trajectoire est réduite à un point. Deux lacets γ_1 et γ_2, contenus dans un ouvert U du plan $\mathbb{C}$, sont dits *homotopes* si l'on peut construire une application $h : I \times I \to \mathbb{C}$ vérifiant :

$$\begin{cases} h(t, a) = \gamma_1(t) \text{ pour tout } t \in I \\ h(t, b) = \gamma_2(t) \\ h(a, u) = h(b, u) \text{ pour tout } u \in I \end{cases}$$

h est donc une « déformation » continue permettant de « passer » de γ_1 à γ_2.

Un ouvert $U \subset \mathbb{C}$ est dit *simplement connexe* lorsque tout lacet γ de U est homotope dans U à un lacet constant (ou homotope à un point).

Le **théorème de Cauchy** exprime alors que, pour une fonction $f : U \to \mathbb{C}$, holomorphe dans un ouvert U simplement connexe, l'intégrale de f le long de tout lacet γ de U est nulle :

$$\int_\gamma f(z)\, dz = 0.$$

Plus généralement, si γ_1 et γ_2 sont deux lacets homotopes d'un ouvert $U \subset \mathbb{C}$ quelconque, pour toute fonction $f : U \to \mathbb{C}$, holomorphe dans U, on a :

$$\int_{\gamma_1} f(z)\, dz = \int_{\gamma_2} f(z)\, dz.$$

Cette intégrale ne dépend donc que de la classe d'homotopie du contour d'intégration (voir *Courbes et surfaces*).

La conséquence est que toute fonction holomorphe dans un ouvert y admet localement une primitive qui est aussi holomorphe ; c'est-à-dire que, pour tout point de cet ouvert, on peut déterminer un voisinage de ce point dans lequel la fonction admet une primitive, qui est holomorphe.

Ce théorème permet d'établir la *formule intégrale de Cauchy*. Celle-ci montre que, à l'intérieur (dans un sens qu'il convient de préciser) d'un chemin fermé de $\mathbb{C}$, le comportement d'une fonction holomorphe (dans un ouvert contenant ce chemin) est déterminé par son comportement le long du chemin, uniquement.

L'indice d'un point z_0 par rapport à une courbe fermée γ se définit par la formule :

$$I(\gamma, z_0) = \frac{1}{2\pi i} \int_\gamma \frac{dz}{z - z_0}$$

et l'on peut montrer qu'il s'agit d'un entier relatif, qui n'est pas modifié lorsque le chemin γ se déforme continûment, mais sans passer par z_0. De plus, dans chaque composante connexe du complémentaire de l'image de γ, le point z_0 peut se déplacer sans altérer la valeur de $I(\gamma, z_0)$.

Un calcul simple dans le cas où γ est un cercle permet alors de conclure de tout ceci que le nombre $I(\gamma, z_0)$ représente en fait la notion intuitive du nombre de fois où la courbe γ « tourne autour » du point z_0 (*fig.* 40).

C'est en ce sens que la notion d'un point intérieur à un chemin fermé se trouve précisée, et la formule de Cauchy s'écrit alors pour un chemin fermé γ d'un ouvert simplement connexe $U \subset \mathbb{C}$, une fonction f holomorphe dans U et un point $a \in U$ tel que γ ne passe pas par a :

$$I(\gamma, a) \cdot f(a) = \frac{1}{2\pi i} \int_\gamma \frac{f(z)\, dz}{z - a}$$

Il existe une réciproque au théorème de Cauchy, le **théorème de Morera** : toute fonction continue $f : D \to \mathbb{C}$, où D est un ouvert de $\mathbb{C}$, telle que

$$\int_\gamma f(z)\, dz = 0$$

— si γ est un *rectangle* fermé contenu dans D — est holomorphe dans D.

Mais la théorie de Cauchy permet surtout, à ce point, de montrer que toute fonction holomorphe dans un ouvert y est analytique. En effet, il suffit de considérer la formule de Cauchy, et de remarquer que l'on a :

$$\frac{1}{z - a} = \frac{1}{z}\left(1 + \frac{a}{z} + \left(\frac{a}{z}\right)^2 + ... + \left(\frac{a}{z}\right)^n + ...\right) \text{ si } |a| < |z|$$

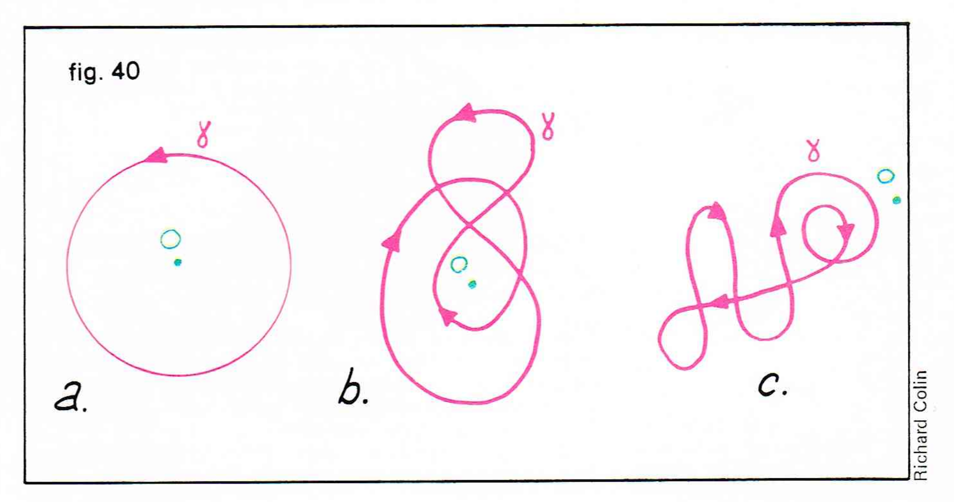
donc on peut écrire :

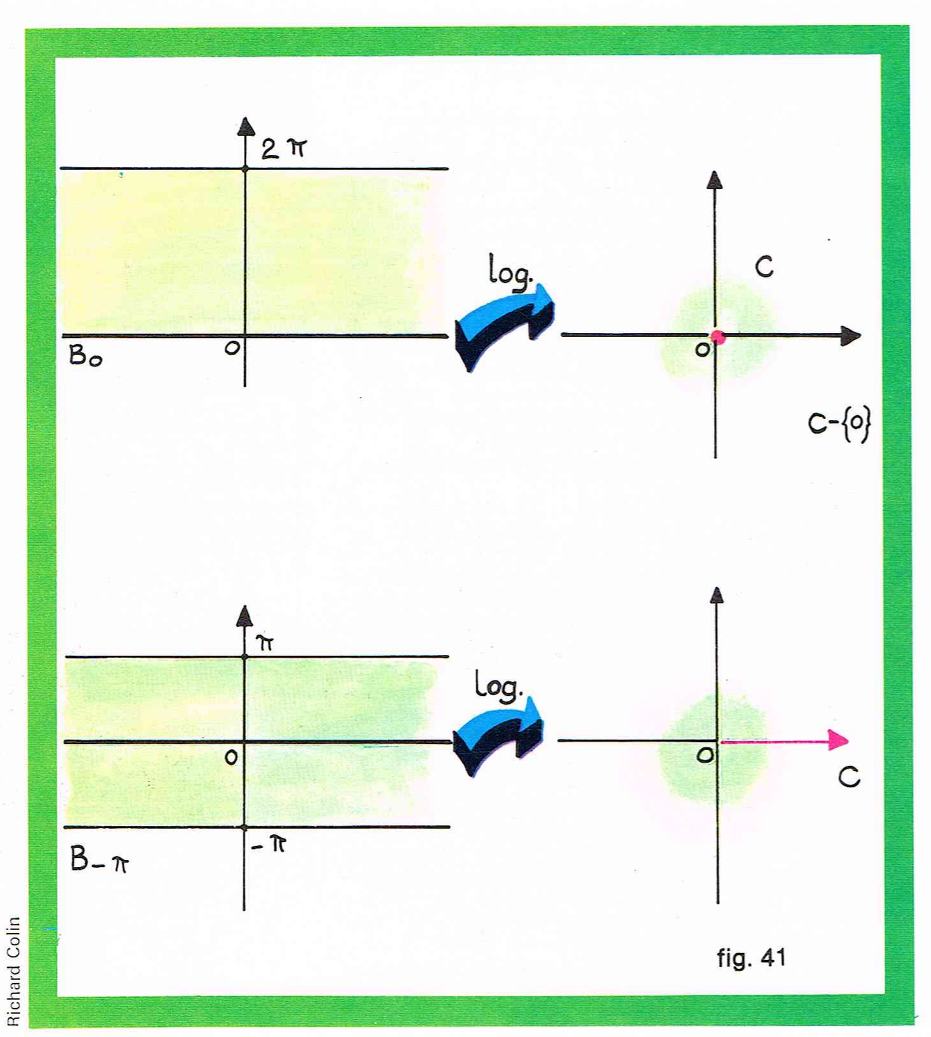
$$f(a) = \frac{1}{2\pi i} \int_\gamma \sum_{n=0}^{\infty} a^n \frac{f(z)}{z^{n+1}}\, dz$$


F. Arborio Mella

▲ ***Le mathématicien français* Augustin-Louis Cauchy *(1789-1857) est l'auteur de méthodes de calcul rigoureuses employées encore de nos jours.***

▼ ***Figure 40 : indice d'un point par rapport à une courbe γ; il est donné par le nombre de fois où la courbe tourne autour du point; a, $I(0, \alpha) = 1$; b, $I(0, \alpha) = 2$; c, $I(0, \alpha) = 0$.***





▲ ***Figure 41 :*** ***représentation de la fonction logarithme complexe (voir développement dans le texte).***

L'interversion des opérations intégration et sommation est possible, car la série converge normalement dans le domaine fermé défini par γ, et l'on obtient bien :

$$f(a) = \sum_{n=0}^{\infty} c_n a^n \quad \text{où} \quad c_n = \frac{1}{2\pi i}\int_{\gamma} \frac{f(z)}{z^{n+1}}\,dz$$

Toute fonction analytique admet une dérivée, donc, pour les fonctions d'une variable complexe les propriétés d'holomorphie et d'analycité sont identiques. Ce résultat, considérable, implique donc qu'une fonction holomorphe est indéfiniment dérivable, donc que la dérivée d'une fonction holomorphe est encore holomorphe. Il est donc légitime de penser que les propriétés des fonctions holomorphes sont très étendues ; on peut citer parmi les plus fondamentales celle que donne le **théorème de Liouville :** toute fonction entière (c'est-à-dire holomorphe dans $\mathbb{C}$ tout entier) ne peut être bornée que si elle est constante.

Ce résultat est obtenu à partir des « *inégalités de Cauchy* » qui précisent le comportement des coefficients de la série de Taylor d'une fonction *f*, holomorphe dans un disque de centre O et de rayon *r* :

$$|c_n| \leqslant \frac{M(r)}{r^n} \quad (n = 0, 1, 2, \ldots)$$

si on pose : $M = \sup_{|z|=r} |f(z)|$

Il faut remarquer, d'ailleurs, qu'un résultat connu sous le nom de « *principe du maximum* » entraîne que, si $M = \sup_{|z|=r} |f(z)|$, alors on a aussi $M = \sup_{|z|\leqslant r} |f(z)|$; c'est-à-dire que, si une fonction holomorphe admet un maximum relatif (au sens des modules) à l'intérieur d'un disque fermé contenu dans le domaine d'holomorphie, alors elle est constante dans ce domaine, pourvu que celui-ci soit un ouvert connexe.

Ces trois résultats précisent encore la différence entre les fonctions d'une variable réelle et les fonctions d'une variable complexe, tout au moins en ce qui concerne la condition de fonction dérivable.

Signalons, de plus, que le théorème de Liouville permet une démonstration très simple du théorème fondamental de l'algèbre, le **théorème de D'Alembert** (voir *Équations algébriques*) ; en effet, si un polynôme P (z) — qui est une fonction entière — n'admettait aucun zéro, alors $\frac{1}{P}$ serait aussi une fonction entière, et puisque $\lim_{z\to\infty} |P(z)| = +\infty$, on aurait $\lim_{z\to\infty} \left|\frac{1}{P(z)}\right| = 0$. La fonction $\frac{1}{P}$ entière et bornée serait donc constante. Il s'ensuit donc par l'absurde que le polynôme P admet un zéro, donc *n* s'il est de degré *n*, en divisant successivement par les monômes $(z - a_k)$, a_k désignant les zéros successifs.

Sans quitter ce sujet, il faut mentionner un résultat qui indique le comportement d'une fonction holomorphe au voisinage d'un zéro quelconque et qui est très étroitement lié au caractère d'analycité de la fonction. Très exactement, si *f* désigne une fonction holomorphe dans un ouvert $\Omega \subset \mathbb{C}$ et $a \in \Omega$ un zéro de *f*, alors, ou bien *f* est identiquement nulle dans un voisinage de *a*, ou bien *a* est le seul zéro de *f* dans un voisinage de *a* assez petit (ce théorème est connu sous le nom de **principe des zéros isolés**). Il ne peut donc exister de point limite d'une suite de zéros de *f* sans que *f* s'annule partout dans un disque ayant ce point pour centre.

Certaines des propriétés ci-dessus ont un caractère « local » ; il faut noter que, dans le passage à un résultat plus « global », c'est la condition de connexité qui est l'hypothèse nécessaire. Ainsi, on peut montrer que, si $\Omega \subset \mathbb{C}$ est un ouvert *connexe*, l'ensemble des zéros d'une fonction holomorphe dans Ω est un ensemble discret (dont tous les points sont isolés) ou bien Ω tout entier.

On montre encore que, si *f* et *g* sont deux fonctions holomorphes dans un ouvert $\Omega \subset \mathbb{C}$, égales sur un sous-ensemble $E \subset \Omega$, non discret, alors elles sont égales partout sur Ω.

Logarithme complexe

On sait qu'on peut définir, dans le cadre des fonctions de variable réelle, la fonction logarithme comme la réciproque de l'application définie sur $\mathbb{R}$ et à valeurs dans $\mathbb{R}_+^*$, $x \mapsto \exp(x)$, puisque, pour ces deux ensembles de départ et d'arrivée, on a une bijection.

L'extension au cas des fonctions de variable complexe est un peu plus délicate.

En posant, pour tout nombre complexe $z = x + iy$, $e^z = e^x \cdot e^{iy} = e^x (\cos y + i \sin y)$, on définit une application $z \mapsto \exp(z)$ sur $\mathbb{C}$ tout entier. Comme $\exp(2\pi i) = 1$, on a $e^{x+iy} = e^{x+i(y+2\pi)}$; donc, on a « périodicité par rapport à la partie imaginaire, de période 2π ». Il suffit donc de partager le plan complexe en bandes horizontales de largeur 2π pour définir une bijection de l'une de ces bandes sur l'ensemble $\mathbb{C} - \{0\}$ par l'application exponentielle complexe ; par exemple, $z \mapsto \exp(z) = e^z$ définit une bijection de la bande $B_0 = \{z \in \mathbb{C} \text{ tels que } 0 < Imz \leqslant 2\pi\}$ sur le plan privé de l'origine, ou bien de la bande ouverte

$$B_{-\pi} \{z \in \mathbb{C} \text{ tels que } |Imz| < \pi\}$$

sur le sous-ensemble

$$\mathbb{C} - \{z \in \mathbb{C} \text{ tels que } Imz = 0 \text{ et } Rez \leqslant 0\},$$

soit $\mathbb{C}$ privé du demi-axe réel négatif (*fig.* 41).

On appelle *détermination principale* du logarithme de *z* la fonction réciproque de l'application $z \mapsto e^z$ de $B_{-\pi}$ sur $\mathbb{C}$ — {demi-axe réel négatif}. On écrira donc

$$\log z = \log |z| + i \arg(z).$$

D'autres déterminations du logarithme sont obtenues en partant d'une autre bande B_α, $\alpha \neq -\pi$. Si l'on note D_α le complémentaire, dans $\mathbb{C}$, de l'image par e^z de la droite $Im\, z = \alpha$, alors e^z définit une bijection de B_α sur D_α, continue ; son application réciproque est appelée $\log_\alpha z$, et vérifie : $\begin{cases} \alpha < I_m(\log z) < \alpha + 2\pi. \\ e^{\log_\alpha z} = z \end{cases}$

Séries de Laurent et résidus

Les séries de Laurent sont un outil qui permet de généraliser le développement en série entière lorsque l'on est en présence d'une fonction holomorphe dans un ouvert, sauf en un nombre fini de points.

Pour une fonction $f : \Omega \to \mathbb{C}$ holomorphe dans Ω, sauf au point $a \in \Omega$, on va d'abord étendre la formule de Cauchy, puis trouver un développement en série qui ne sera plus celui de f dans un disque, mais dans une couronne. Une telle extension est faite pour des *points isolés* de Ω, c'est-à-dire des points frontières tels qu'on puisse trouver pour chacun un voisinage ne contenant aucun autre point frontière. Si a désigne un point isolé de Ω, on cherchera donc à caractériser f dans un disque : $0 < |z - a| < r$ (privé de son centre a) ; pour plus de généralité, on considérera même une couronne

$$r_1 < |z - a| < r_2.$$

La difficulté qui surgit alors est qu'un tel ensemble n'est pas simplement connexe. Or, cependant, les deux lacets γ_1 et γ_2 définis par les cercles $|z - a| = r_1$ et $|z - a| = r_2$ sont homotopes dans Ω', et cela permet donc d'écrire que, pour toute fonction g holomorphe dans Ω', si

$$\Omega' = \Omega - \{a\} : \int_{\gamma_1} g(z)\, dz = \int_{\gamma_2} g(z)\, dz \quad (\textit{fig. } 42).$$

En prenant la fonction :

$$\begin{cases} g(z) = \dfrac{f(t) - f(z)}{t - z} & \text{lorsque } t \neq z \\ g(z) = f'(z) & \text{lorsque } t = z \end{cases}$$

qui est bien holomorphe dans la couronne

$$C_a = \{z \in \mathbb{C} \text{ tels que } r_1 < |z - a| < r_2\}$$

on obtient :

$$f(z) = \frac{1}{2\pi i} \int_{\gamma_2} \frac{f(t)}{t - z}\, dt - \frac{1}{2\pi i} \int_{\gamma_1} \frac{f(t)}{t - z}\, dt$$

qui généralise la formule de Cauchy du paragraphe précédent.

Un développement en série de $\dfrac{1}{t - z}$ pour $|t| = r_2$ uniformément convergent dans C_a est $\dfrac{1}{t}\left[\sum_0^{\infty} \left(\dfrac{z}{t}\right)^n\right]$. De même, un développement en série de $\dfrac{1}{t - z}$, mais pour $|t| = r_1$, uniformément convergent dans C_a, est : $-\dfrac{1}{z}\left[\sum_0^{\infty} \left(\dfrac{t}{z}\right)^n\right]$.

Ceci permet alors d'obtenir le développement dit « *en série de Laurent* » ·

$$f(z) = \sum_{n=-\infty}^{+\infty} c_n (z - a)^n = \sum_{n=-\infty}^{-1} c_n (z - a)^n + \sum_{n=0}^{\infty} c_n (z - a)^n = \sum_{n=1}^{\infty} \frac{c_{-n}}{(z - a)^n} + \sum_{n=0}^{\infty} c_n \cdot (z - a)^n$$

où l'on a :

$$c_n = \frac{1}{2\pi i} \int_{\gamma} \frac{f(z)}{(z - a)^{n+1}}\, dz$$

γ désignant un cercle centré en a, et de rayon r, $r_1 < r < r_2$

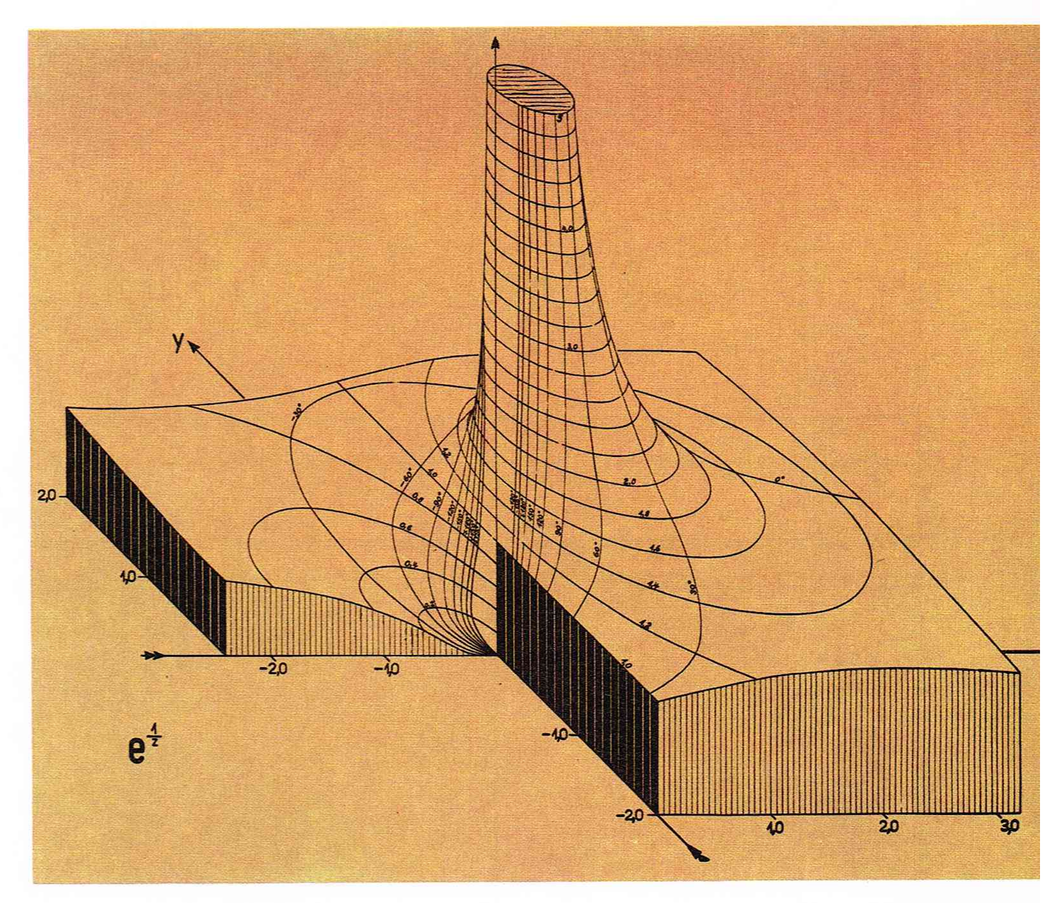
On peut montrer que la série $\sum_{n=1}^{\infty} \dfrac{c_{-n}}{(z - n)^n}$ est convergente pour z tel que $|z - a| > r_1$, tandis que la série $\sum_{n=1}^{\infty} c_n (z - a)^n$ l'est pour z tel que $|z - a| < r_2$.

Il faut noter encore que le développement en série de Laurent, une série en puissances de $(z - a)$ et une autre en puissances de $\dfrac{1}{(z - a)}$, est unique dans la couronne C_a.

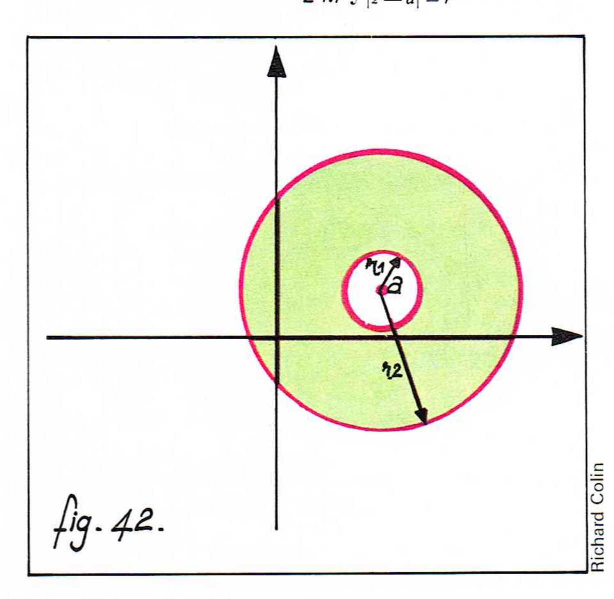
Bien entendu, lorsque f est holomorphe dans Ω tout entier, on retrouve le développement de Taylor, puisque $c_{-n} = 0$; le point a est dit *point régulier*. Sinon, le point a est dit *point singulier;* c'est très exactement un pôle d'ordre q lorsque $c_{-n} = 0$ pour $n > q$ avec $c_{-q} \neq 0$ (ainsi, l'origine 0 est un pôle d'ordre 37 pour la fonction $\dfrac{e^z}{z^{37}}$) et c'est un *point singulier essentiel* dans les autres cas : ainsi l'origine 0 est un point singulier essentiel pour la fonction $\dfrac{e^z}{z}$.

Pour tout point singulier (pôle ou point singulier essentiel), le coefficient c_{-1} de la série de Laurent de f porte le nom de *résidu de f en a* que l'on note Rés (f, a) :

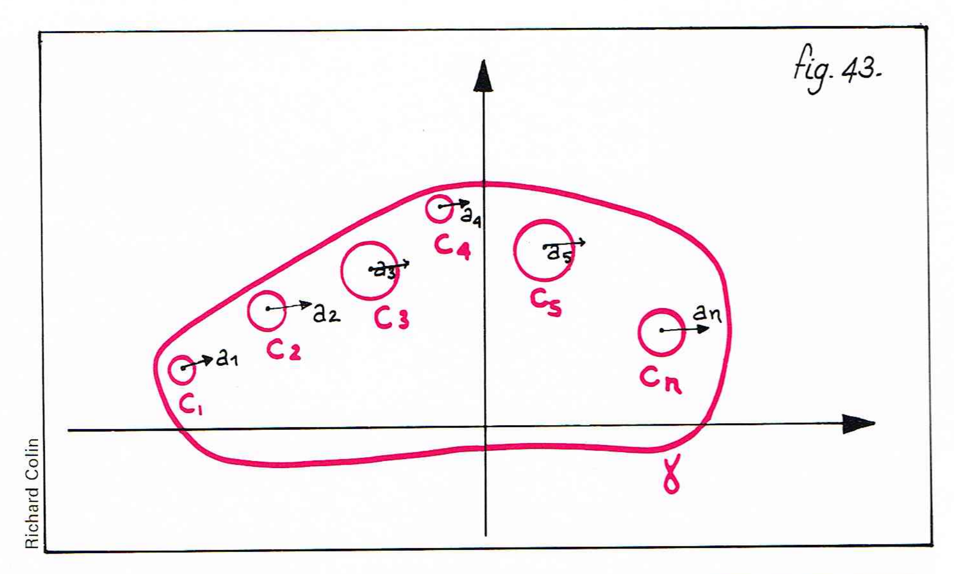
$$\text{Rés}(f, a) = c_{-1} = \frac{1}{2\pi i} \int_{|z-a|=r} f(z)\, dz.$$

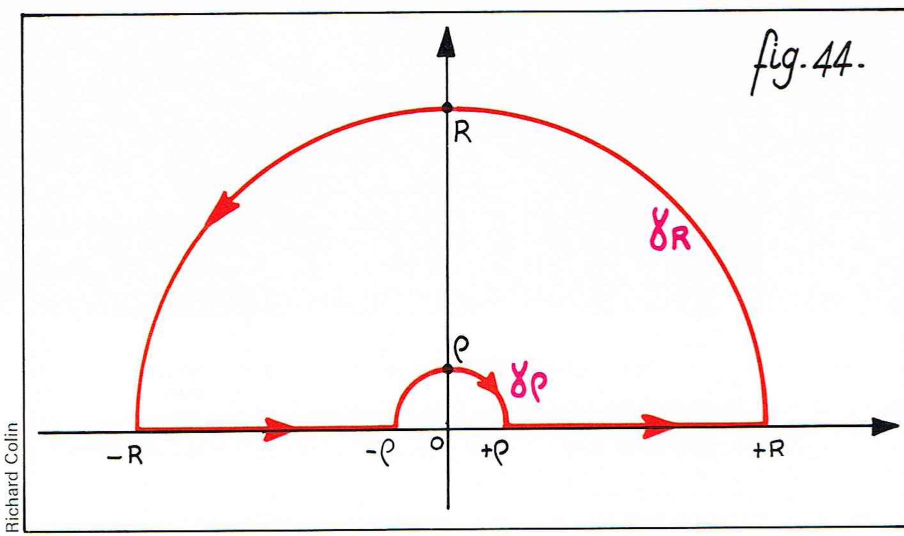


▲ ***Point singulier de la fonction complexe*** $e^{\frac{1}{z}}$



◄ ***Figure 42 : couronne*** C_a ***définie par*** $r_1 < |z - a| < r_2$.

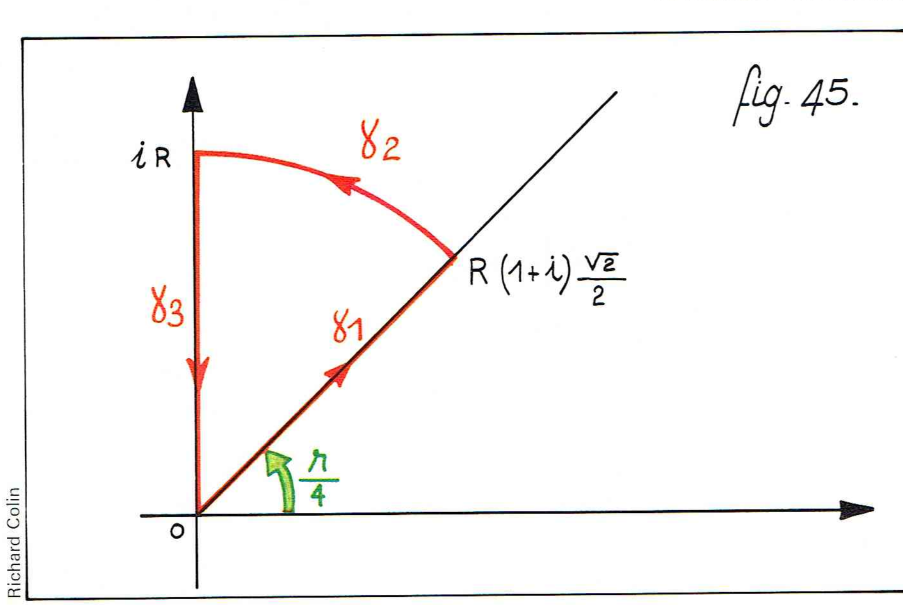




▲ *En haut, figure 43 : théorème des résidus.*

Ci-dessus, figure 44 : voir développement dans le texte-ci-contre.

▼ *Figure 45 : l'intégration de la fonction $z \to e^{z^2}$ sur le contour γ donne 0 grâce au théorème de Cauchy.*



On peut alors transcrire le théorème de Cauchy pour une fonction ayant un nombre fini de points singuliers, et obtenir le **théorème des résidus** : *Pour une fonction* f : $\Omega \to \mathbb{C}$, *holomorphe dans* Ω *sauf en un nombre fini de points singuliers* a_n (n = 1, 2, ..., *N*), *si* γ *est une courbe fermée contenue dans* Ω, *ne contenant aucun des points* a_n, *on a :*

$$\frac{1}{2\pi i}\int_\gamma f(z)\,dz = \sum_{n=1}^{N} \mathrm{I}(\gamma, a_n) \cdot \text{Rés}(f, a_n).$$

Cette formule, qui est bien évidemment une extension de la propriété des fonctions holomorphes, s'explique en remarquant qu'elle donne la limite de l'action des points singuliers sur la valeur de l'intégrale. Plus exactement, si c_n désigne le cercle de centre a_n et de rayon r_n (choisi pour que les cercles soient disjoints et ne coupent pas γ), on a $\mathrm{I}(c_n, a_n) = 0$ si $n \neq m$ et $\mathrm{I}(c_n, a_n) = 1$ (*fig.* 43).

Si l'on pose : $\gamma' = \gamma + \alpha_1 c_1 + ... + \alpha_n c_n$ avec $\alpha_n = -\mathrm{I}(\gamma, a_n)$, on a alors $\mathrm{I}(\gamma', a_n) = 0$ pour $n = 1, 2, ..., \mathrm{N}$. Ce qui permet d'écrire que :

$$0 = \frac{1}{2\pi i}\int_{\gamma'} f(z)\,dz =$$

$$\frac{1}{2\pi i}\left\{\int_\gamma f(z)\,dz - \sum_{n=1}^{n=N} \mathrm{I}(\gamma, a_n)\int_{c_n} f(z)\,dz\right\}$$

et il ne suffit plus que d'appliquer la définition du résidu de f au point a_n.

Des conditions restrictives sur Ω, délicates à expliquer, doivent en fait être rajoutées aux hypothèses du théorème ; elles sont satisfaites dès lors que Ω est un ouvert borné et simplement connexe.

Le théorème des résidus est l'un des outils les plus connus pour les calculs d'intégrales impropres (il est même parfois utilisé dans la résolution approchée de certaines équations).

En voici quelques exemples :

— Pour le calcul de $\int_{-\infty}^{+\infty} \frac{\sin x}{x}\,dx$, dont on montre aisément la convergence en intégrant par parties

$$f(x) = \frac{\sin x}{x}$$

sur [0, 1] après avoir posé $f(0) = 1$, on a recours à l'intégrale $\int_\gamma \frac{e^{iz}}{z}\,dz$ où $\gamma = \gamma_R - \gamma_\rho$ avec :

$$\begin{cases} \gamma_R \text{ paramétré par } Re^{it} \\ \gamma_\rho \text{ paramétré par } \rho e^{it} \text{ pour } t \in [0, \pi] \end{cases}$$ (*fig.* 44).

On applique le théorème de Cauchy :

$$0 = \int_\gamma \frac{e^{iz}}{z}\,dz = \int_{-R}^{-\rho} \frac{e^{ix}}{x}\,dx - \int_0^\pi e^{i\rho e^{it}} \cdot i\,dt + \int_\rho^R \frac{e^{ix}}{x}\,dx + \int_0^\pi e^{iRe^{it}} \cdot i\,dt$$

en décomposant sur chaque « partie » de γ.

Puisque

$$I_m\left[\int_{-R}^{-\rho} \frac{e^{ix}}{x}\,dx + \int_\rho^R \frac{e^{ix}}{x}\,dx\right] = 2\int_\rho^R \frac{\sin x}{x}\,dx$$

et que :

$$\lim_{\rho \to 0} \int_0^\pi e^{i\rho e^{it}} \cdot i\,dt = i\pi$$

ainsi que $\lim_{R \to \infty} \int_0^\pi e^{iRe^{it}} \cdot i\,dt = 0$

que l'on vérifie aisément, on en déduit que :

$$\int_{-\infty}^{+\infty} \frac{\sin x}{x}\,dx = \pi$$

[La notation *Im* (A) désigne la partie imaginaire du nombre complexe A.]

— L'intégration de la fonction $z \to e^{z^2}$ sur le contour γ de la figure 45 donne 0 grâce au théorème de Cauchy. La décomposition de $\int_\gamma e^{z^2}\,dz$ selon chacun des trois

« morceaux » de γ, en tenant compte du paramétrage :

$$\begin{cases} \text{sur } \gamma_1 : z = re^{i\frac{\pi}{4}} & \text{avec } 0 \leqslant r \leqslant \text{R}. \\ \text{sur } \gamma_2 : z = \text{R}e^{i\theta} & \text{avec } \frac{\pi}{4} \leqslant \theta \leqslant \frac{\pi}{2}. \\ \text{sur } \gamma_3 : z = ir & \text{avec } 0 \leqslant r \leqslant \text{R}. \end{cases}$$

permet à l'aide de majorations classiques de calculer les célèbres *intégrales de Fresnel :*

$$\int_0^\infty \cos u^2 du \quad \text{et} \quad \int_0^\infty \sin u^2 du$$

qui sont égales toutes deux à $\frac{\sqrt{2\pi}}{4}$.

Dans ces deux exemples, on ne fait usage que du théorème de Cauchy, car les fonctions considérées le permettaient. Les exemples suivants donnent des applications de la théorie des résidus.

— L'intégrale $\int_{-\infty}^{+\infty} \frac{dx}{(1+x^2)^2}$ se calcule en intégrant la fonction $\frac{1}{(1+z^2)^2}$ sur un demi-cercle, situé dans le demi-plan supérieur centré à l'origine et dont on fait croître indéfiniment le rayon. Cette fonction admet un pôle au point $+i$ dans ce domaine. Pour déterminer le résidu en ce point, il suffit de faire un changement de variable $z = t + i$ et de chercher un développement limité de la fonction $\frac{1}{(1+z^2)^2} = \frac{1}{t^2(t+2i)^2}$ au voisinage du point $t = 0$.

Par décomposition simple,

$$\frac{1}{t^2(t+2i)^2} = \frac{\text{A}}{t^2} + \frac{\text{B}}{t} + \mathcal{O}(t),$$

ce qui donne :

$$\frac{1}{t^2 \cdot (t+2i)^2} = \frac{1}{t^2} \cdot \frac{-1}{4} \left(\frac{1}{\frac{t}{2i} + 1} \right)^2 =$$

$$\frac{1}{t^2} \cdot \frac{-1}{4} \left(1 - \frac{t}{2i} \right) + \mathcal{O}(t)$$

Dans cette expression, le coefficient de $\frac{1}{t}$, égal à $\frac{1}{4i}$, est donc le résidu recherché.

Par conséquent $\int_\gamma \frac{dz}{(1+z^2)^2} = 2\pi i \cdot \frac{1}{4i} = \frac{\pi}{2}$. Il suffit (*fig.* 46) de décomposer la courbe d'intégration en le segment [— R, + R] et le demi-cercle, puis de remarquer que, lorsque $\text{R} \to \infty$, l'intégrale sur le demi-cercle tend vers zéro en module, pour conclure que :

$$\int_{-\infty}^{+\infty} \frac{dx}{(1+x^2)^2} = \frac{\pi}{2}.$$

— Pour calculer $\int_0^{+\infty} \frac{(\text{Log } x)^2}{1+x^2} dx$, on considère la fonction : $f(z) = \frac{(\text{Log } z)^2}{1+z^2}$ et le contour de la figure 47.

Le paramétrage selon les diverses composantes de ce contour donne :

$$\frac{(\text{Log } x)^2}{1+x^2} \quad \text{sur } [\varepsilon, \text{R}]$$

$$\frac{(\text{Log } x + i\pi)^2}{1+x^2} \quad \text{sur } [-\text{R}, -\varepsilon]$$

ainsi que :

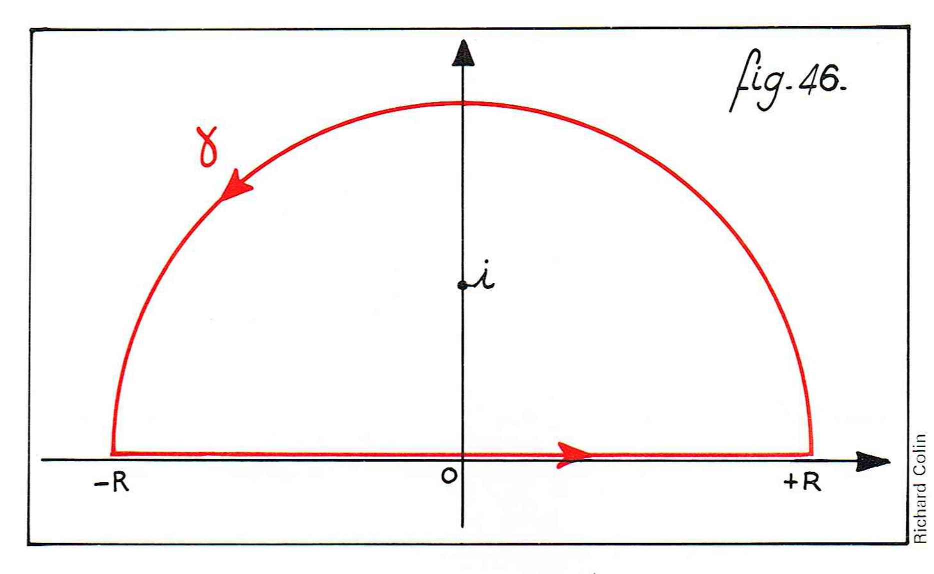
$$\frac{(\text{Log R} + i\theta)^2}{1+\text{R}^2 e^{2i\theta}} \quad \text{sur } \gamma_\text{R}$$

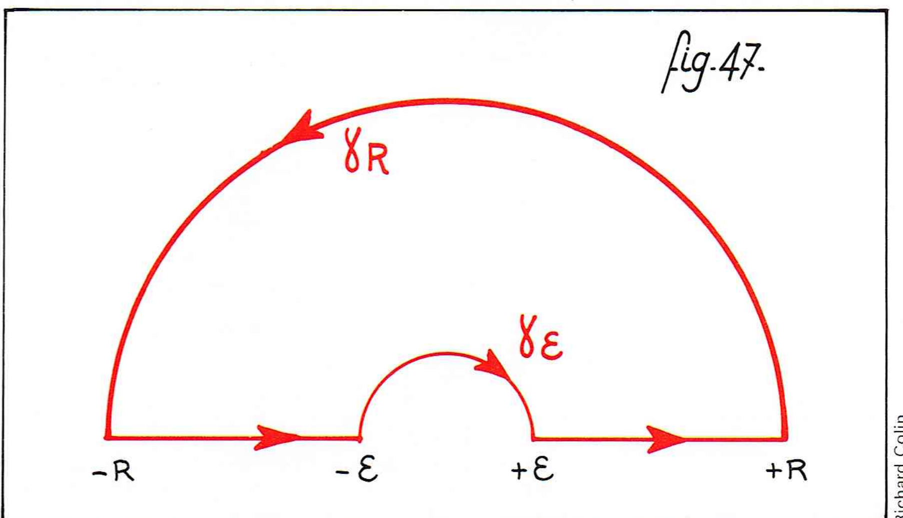
$$\frac{(\text{Log } \varepsilon + i\theta)^2}{1+\varepsilon^2 e^{2i\theta}} \quad \text{sur } \gamma_\varepsilon$$

Puisque $\frac{(\text{Log R} + i\theta)^2}{1+\text{R}^2 e^{2i\theta}} = \mathcal{O}\,\frac{[(\text{Log R})^2]}{\text{R}^2}$,

il s'ensuit que : $\lim\limits_{\text{R}\to\infty} \int_{\gamma_\text{R}} f(z)\, dz = 0$,

de même : $\lim\limits_{\varepsilon\to 0} \int_{\gamma_\varepsilon} f(z)\, dz = 0$.





▲ En haut, figure 46 : décomposition de la courbe d'intégration en le segment [— R, + R] et le demi-cercle. Ci-dessus, figure 47 : voir développement de cette application de la théorie des résidus dans le texte.

Le théorème des résidus montre donc que :

$$\lim_{\substack{\varepsilon \to 0 \\ \text{R}\to\infty}} \left[\int_\varepsilon^\text{R} \frac{(\text{Log } r)^2 + (\text{Log } r + i\pi)^2}{1+r^2} dr \right] = 2\pi i \, \Sigma \, \text{Rés}\,(f, z_k)$$

la sommation concernant tous les pôles z_k du demi-plan supérieur. En fait, il n'y en a qu'un seul : $z = i$. On montre que, pour une fraction $\frac{\text{A}(z)}{\text{B}(z)}$, le résidu en un pôle simple z_k vaut exactement $\frac{\text{A}(z_k)}{\text{B}'(z_k)}$. Cela donne ici :

$$\text{Rés}\,(f, i) = \frac{(\text{Log } i)^2}{2i} = \frac{-\pi^2}{8i}.$$

La suite de la décomposition de l'intégrale est :

$$\lim_{\substack{\varepsilon \to 0 \\ \text{R}\to\infty}} \int_\gamma f(z)\, dz = \int_0^\infty \frac{2(\text{Log } r)^2}{1+r^2} dr + 2\pi i \int_0^\infty \frac{\text{Log } r}{1+r^2} dr - \pi^2 \int_0^\infty \frac{dr}{1+r^2} = -\frac{\pi^2}{8i} \cdot 2\pi i = -\frac{\pi^3}{4}.$$

Les résultats classiques :

$$\int_0^\infty \frac{\text{Log } r}{1+r^2} dr = 0 \quad \text{et} \quad \int_0^\infty \frac{dr}{1+r^2} = \frac{\pi}{2}$$

permettent alors d'écrire le résultat :

$$\int_0^\infty \frac{(\text{Log } x)^2}{1+x^2} dx = \frac{\pi^3}{8}$$

Prolongement analytique

L'idée simple du prolongement par continuité pour les fonctions réelles de variable réelle, que l'on peut illustrer par l'exemple de l'application définie sur $\mathbb{R} - \{0\}$, $f : x \mapsto \exp\left(-\frac{1}{x^2}\right)$, pour laquelle on peut définir une fonction continue sur $\mathbb{R}$, égale à f sur $\mathbb{R} - \{0\}$, soit :

$$g \begin{cases} x \mapsto \exp\left(-\frac{1}{x^2}\right) & \text{si } x \neq 0 \\ 0 \mapsto 0 \end{cases}$$

puisque $\lim_{x \to 0} \exp\left(-\frac{1}{x^2}\right) = 0$ se transcrit pour les fonctions de variable complexe par une notion plus forte : celle du *prolongement analytique*. Partant d'un ouvert $\Omega \subset \mathbb{C}$ et d'une fonction $f : \Omega \to \mathbb{C}$, holomorphe dans Ω, on cherche à construire un ouvert $\Omega' \supset \Omega$, une fonction $g : \Omega' \to \mathbb{C}$ holomorphe dans Ω' telle que $f(z) = g(z)$ si $z \in \Omega$.

En notant R le rayon de convergence (non nul et fini) de la série entière égale à f dans Ω, le problème du prolongement se pose sur la frontière du disque de convergence, c'est-à-dire pour des points $t \in \mathbb{C}$ tels que $|t| = \mathrm{R}$. Le point t est alors *point régulier* si le prolongement est possible dans un voisinage de t; sinon, c'est un *point singulier*. Par exemple, grâce aux résultats obtenus sur les séries on sait que $t = +1$ est un point singulier pour $\sum^{\infty} z^n$, que $t = +1$ et $t = -1$ sont points singuliers pour $\sum^{\infty} z^{2n}$, que $t = +1$ est point singulier pour $\sum^{\infty} \frac{z^n}{n}$.

La caractérisation de ces points peut se faire au moyen du **théorème** suivant : soit t un point du disque de convergence — de rayon R, centré à l'origine — de f, tel que $|t| = \mathrm{R}$. Soit R_a le rayon de convergence de la série $\sum^{\infty} c_n(a)(z-a)^n$ qui représente f au voisinage de $a = r \cdot t$ (où $0 < r < 1$).

Alors : t est régulier $\Leftrightarrow \mathrm{R}_a > \mathrm{R} - |a|$

t est singulier $\Leftrightarrow \mathrm{R}_a = \mathrm{R} - |a|$ (*fig.* 48).

Il est possible d'affirmer l'existence d'au moins un point singulier sur le cercle de convergence d'une fonction analytique. En particulier, si $\sum^{\infty} c_n z^n$ représente f dans Ω et si les coefficients c_n sont tous réels positifs, le point $t = \mathrm{R}$ est un point singulier.

Développements eulériens — La fonction Γ

On appelle *fonction méromorphe* sur un ouvert Ω du plan complexe une fonction holomorphe sur un ouvert Ω' obtenu en privant Ω de points isolés qui soient des pôles de cette fonction. Ceci revient à dire que, au voisinage de chaque point de Ω, une fonction méromorphe est un quotient de deux fonctions holomorphes.

On doit à Cauchy une méthode pour exprimer une fonction méromorphe dans $\mathbb{C}$ sous forme de série convergente, en dehors des pôles ; il s'agit en fait de généraliser la décomposition des fractions rationnelles en « éléments simples ». Cette méthode permet d'exprimer une fonction impaire, en tout point autre qu'un pôle, par une double sommation dès lors que le module de la valeur de cette fonction sur des cercles de rayons croissant indéfiniment, sans passer par un des pôles, est uniformément borné, soit : $|f(r_n e^{i\rho})| \leqslant \mathrm{K}$ pour tout n et tout ρ.

En désignant par a_k les pôles de f, on a :

$$f(z_0) = \sum_{n=0}^{\infty} \left(\sum_{r_n < |a_k| < r_{n+1}} \text{Rés} \left\{ \frac{f(z)}{z - z_0}, a_k \right\} \right)$$

L'application de cette méthode permet d'obtenir le développement eulérien de cotg z :

$$\text{cotg } z = \frac{1}{z} + 2z \sum_{n=1}^{\infty} \frac{1}{z^2 - n^2\pi^2}$$

puis celui de sin z :

$$\sin z = z \cdot \prod_{n=1}^{\infty} \left(1 - \frac{z^2}{n^2\pi^2}\right)$$

D'une façon générale, le **théorème de factorisation de Weierstrass** permet d'exprimer une fonction entière sous forme de produit infini à partir de la connaissance des points où elle s'annule.

Si $\sum_1^{\infty} \frac{1}{|z_n|^{p+1}} < \infty$, alors, en posant

$$\mathrm{P}_n(z) = \left(\frac{z}{z_n}\right) + \ldots + \frac{1}{p}\left(\frac{z}{z_n}\right)^p$$

le produit infini $\prod_1^{\infty} \left(1 - \frac{z}{z_n}\right) \cdot e^{\mathrm{P}_n(z)}$ converge uniformément sur tout compact et représente une fonction entière dont les zéros sont les points z_n.

C'est comme limite d'un produit infini que la célèbre *fonction gamma*, bien connue des statisticiens, des physiciens, etc., a été l'objet des premiers travaux (Euler, Gauss, Weierstrass).

On pose d'abord :

$$\Pi(x, n) = \frac{n!}{(x+1)(x+2)\ldots(x+n)} \cdot n^x$$

Puis l'on définit :

$$\Gamma(x+1) = \Pi(x) = \lim_{n \to \infty} \Pi(x, n)$$

Si l'on définit la *constante d'Euler* C par :

$$\mathrm{C} = \lim_{n \to \infty} \left(1 + \frac{1}{2} + \ldots + \frac{1}{n} - \text{Log } n\right),$$

on peut alors montrer que :

$$\frac{1}{\Pi(x)} = \frac{1}{\Gamma(x+1)} = e^{\mathrm{C}x} \cdot \prod_1^{\infty} \left[\left(1 + \frac{x}{p}\right) e^{-\frac{x}{p}}\right]$$

et cette fonction est alors définie pour tout x, réel, positif strictement.

Il est alors possible de prolonger sa définition sur tout le plan complexe privé de l'origine :

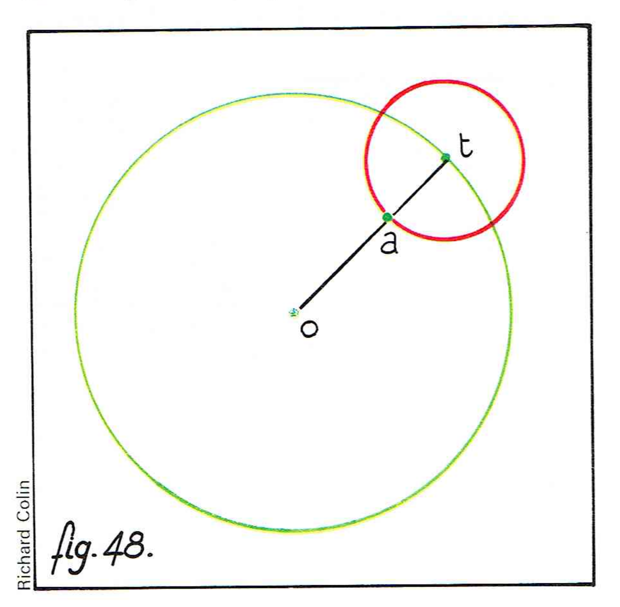
$$\frac{1}{\Gamma(z)} = ze^{\mathrm{C}z} \prod_1^{\infty} \left[\left(1 + \frac{z}{p}\right) e^{-\frac{z}{p}}\right]$$

et la fonction $z \mapsto \Gamma(z)$ obtenue est analytique, sauf aux points : $z = -1, -2, \ldots, -n, \ldots$, qui sont des pôles simples. On peut remarquer qu'elle ne s'annule jamais puisque $\frac{1}{\Gamma(z)}$ ne possède pas de pôle.

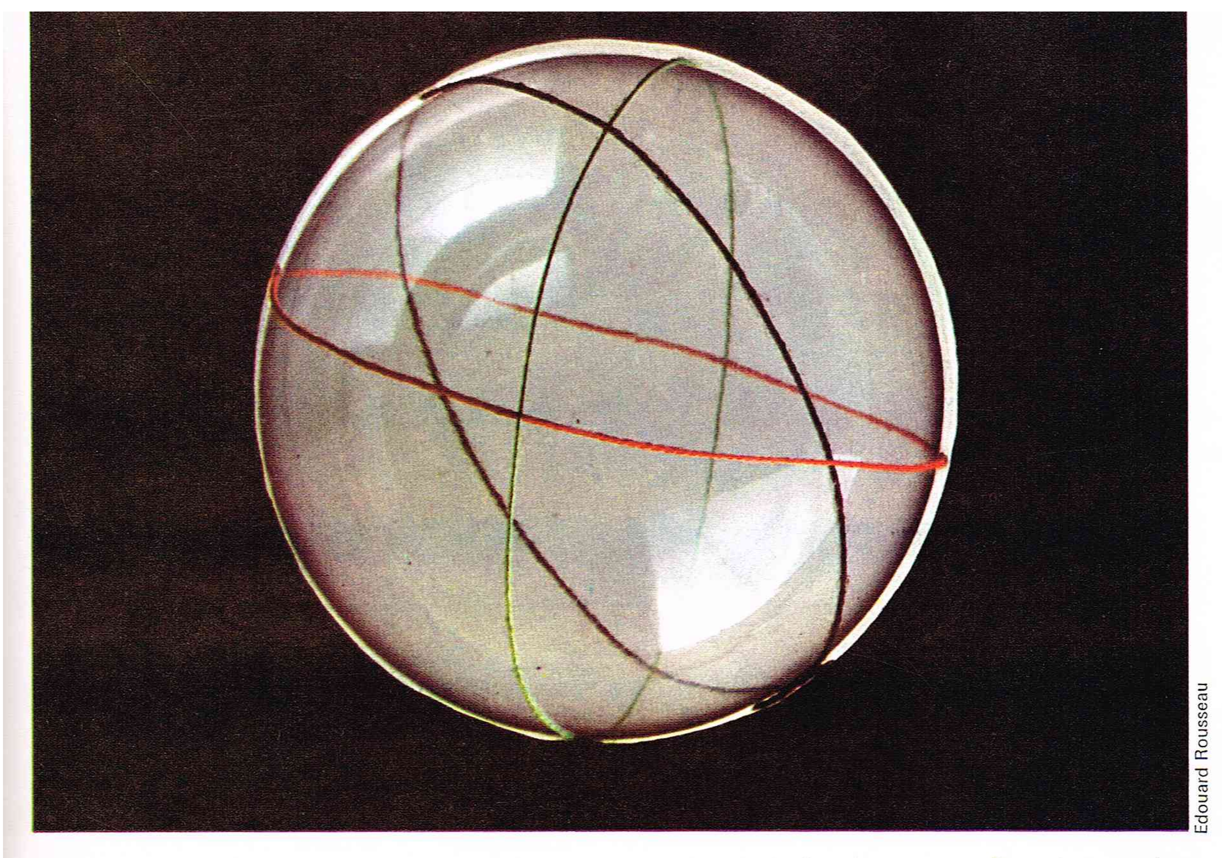
Un calcul simple montre que :

$$\frac{1}{\Gamma(z)} \cdot \frac{1}{\Gamma(-z)} = -z^2 \prod_1^{\infty} \left(1 - \frac{z^2}{p^2}\right) \quad \text{pour } z \notin \mathbb{Z}$$

ce qui, grâce au développement eulérien de sin z, donne



▶ *Figure 48 : points réguliers et points singuliers.*



Édouard Rousseau

◀ **Géométrie de Riemann :** ***la somme des angles d'un triangle n'est pas égale à deux droits.***

la *formule des compléments :* $\Gamma(z)\cdot\Gamma(1-z)=\dfrac{\pi}{\sin \pi z}$ que l'on obtient en notant que : $\Gamma(1-z)=-z\Gamma(-z)$ [cette relation est obtenue à partir de la plus générale : $x\Gamma(x)=\Gamma(x+1)$ que l'on retrouvera plus loin].

On en déduit que : $\Gamma\left(\frac{1}{2}\right)=\sqrt{\pi}$.

Il est ainsi possible de construire des valeurs successives de cette fonction. C'est à Legendre (1752-1833) que l'on doit la présentation de $\Gamma(z)$ sous la forme d'une intégrale définie. Posant : $E(z,n)=\int_0^n\left(1-\frac{t}{n}\right)^n\cdot t^{z-1}\,dt$, il en déduit :

$$E(z,n)=\int_0^1(1-u)^n\cdot u^{z-1}\cdot n^z du = n^z\int_0^1(1-u)^n u^{z-1}\,du,$$

puis : $E(z,n)=\dfrac{n!\,n^z}{z(z+1)\ldots(z+n)}=\dfrac{1}{z}\Pi(z,n)$ en intégrant par parties. Ceci lui permet de redéfinir la fonction $\Gamma(z)$ lorsque $Re(z)>0$ par :

$$\Gamma(z)=\int_0^\infty e^{-t}\cdot t^{z-1}\,dt$$

et d'obtenir ainsi sur le demi-plan ouvert où $Re(z)>0$ une fonction holomorphe.

Une simple intégration par parties permet alors de retrouver la relation : $\Gamma(z+1)=z\Gamma(z)$.

L'étude de l'équation fonctionnelle $\Gamma(z+1)=z\Gamma(z)$ a amené à redéfinir la *fonction gamma* comme étant la solution du système :

$$\begin{cases}\varphi(x+1)=x\varphi(x)\\ \varphi(1)=1\end{cases}$$

soit, en posant $g(x)=\text{Log}\,\varphi(x)$:

$$\begin{cases}g(x+1)-g(x)=\text{Log}(x)\\ g(1)=0\end{cases}$$

Il existe une et une seule fonction *convexe* g vérifiant ces deux dernières relations.

La relation fonctionnelle $\Gamma(x+1)=x\Gamma(x)$ admet évidemment comme conséquence que $\Gamma(n+1)=n!$ pour tout entier n positif. Ceci justifie l'appellation parfois donnée de *fonction « interpolant les factorielles »;* en outre, on retrouve que $0!=1$. Elle permet aussi d'établir la célèbre *formule de Stirling :*

$$n!=\sqrt{2\pi n}\cdot n^n\cdot e^{-n}\left(1+\varepsilon(n)\right) \text{ avec } \lim_{n\to\infty}\varepsilon(n)=0,$$

cas particulier de : $\Gamma(x)=\sqrt{2\pi x}\cdot x^{x-1}\cdot e^{-x}(1+\varepsilon(x))$.

La *fonction bêta* peut être considérée comme une généralisation de la fonction gamma. Pour deux nombres complexes x et y de parties réelles positives, on pose :

$$B(x,y)=\int_0^1 t^{x-1}(1-t)^{y-1}\,dt$$

et l'on montre que $B(x,y)=B(y,x)$.

Un calcul simple donne :

$$B(x,y)=\frac{\Gamma(x)\cdot\Gamma(y)}{\Gamma(x+y)}$$

Cette fonction vérifie les équations fonctionnelles suivantes :

$$B(x,y)=B(x+1,y)+B(x,y+1)=\frac{x+y}{x}B(x+1,y)=\frac{x+y}{y}B(x,y+1)$$

$$B(x+1,y+1)=\frac{xy}{(x+y)(x+y+1)}B(x,y)$$

et permet entre autres, grâce à la formule :

$$B(x,x)=2^{1-2x}\cdot B\left(x,\frac{1}{2}\right)$$

de retrouver la formule de « duplication » :

$$2\sqrt{\pi}\,\Gamma(2x)=2^{2x}\cdot\Gamma(x)\cdot\Gamma\left(x+\frac{1}{2}\right).$$

Très fréquemment utilisées, ces fonctions ont un impact particulier en probabilités.

Représentation conforme

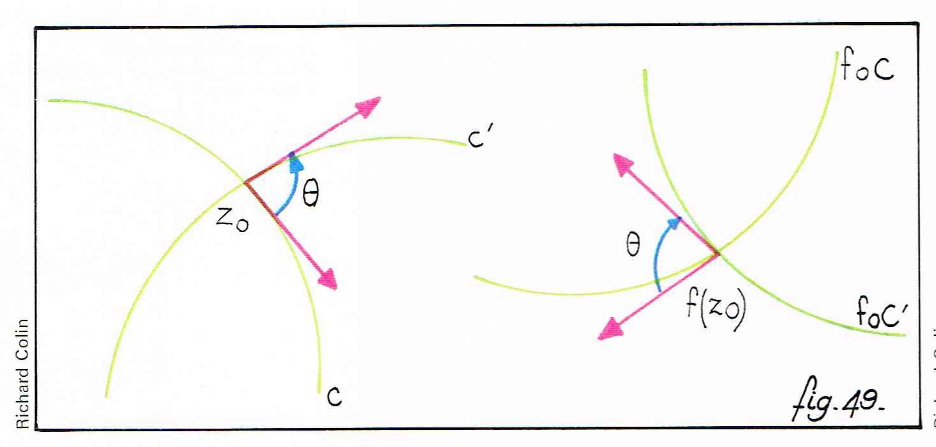
C'est au mathématicien allemand Riemann (1826-1866) que l'on doit la notion de *courbe analytique complexe,* ou encore *surface de Riemann,* qui a conduit au développement actuel de la géométrie algébrique. Reprenant les travaux de Gauss sur les bijections différentiables qui conservent les angles, il leur donna une nouvelle impulsion en leur appliquant les résultats alors connus de la théorie des fonctions de variable complexe.

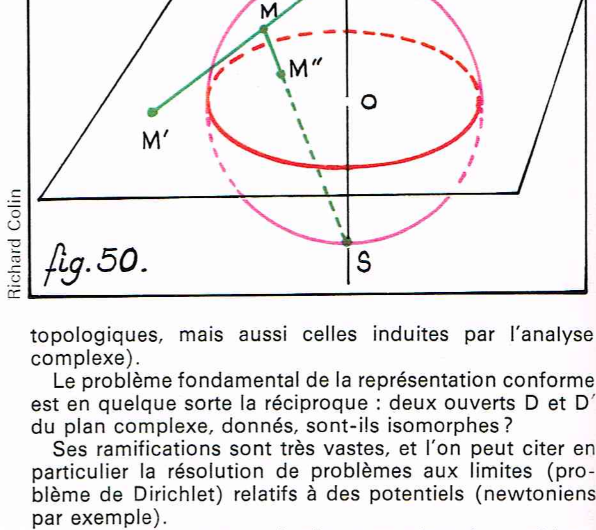
En effet, la notion d'holomorphie recouvre la notion géométrique de similitude : en rapportant l'espace vectoriel réel $\mathbb{C}$ (de dimension 2) à sa base usuelle $\{1, i\}$, une similitude directe de centre O, $z\mapsto u\cdot z$ où u est un nombre complexe non nul, $u=a+ib$, est linéaire et admet relativement à cette base la matrice :

$$\begin{pmatrix}a & -b\\ b & a\end{pmatrix}$$

Si maintenant on pose $u=f'(z_0)$, où f est une fonction holomorphe en z_0 : $f=P+iQ$, les conditions de Cauchy-Riemann :

$$\frac{\partial P}{\partial x}=\frac{\partial Q}{\partial y} \quad\text{et}\quad \frac{\partial P}{\partial y}=-\frac{\partial Q}{\partial x}$$





▲ *A gauche, figure 49 : conservation des angles. A droite, figure 50 : les deux projections stéréographiques.*

montrent bien que la matrice jacobienne de f, soit :

$$\begin{pmatrix} \dfrac{\partial P}{\partial x} & \dfrac{\partial Q}{\partial x} \\ \dfrac{\partial P}{\partial y} & \dfrac{\partial Q}{\partial y} \end{pmatrix}, \text{ est une matrice de similitude.}$$

Par conséquent, si l'on définit une *transformation conforme* comme une application d'un domaine $D \subset \mathbb{R}^2$ dans $\mathbb{R}^2$ différentiable, telle que l'application linéaire dérivée conserve les angles orientés, alors toute fonction f, holomorphe en un point z_0 (resp. sur un domaine D), définit une transformation conforme en z_0 (resp. sur D) dès lors que $f'(z_0) \neq 0$ (resp. $f'(z) \neq 0$ pour tout $z \in D$).

Cette notion d'application conforme introduite par la *conservation des angles* pour l'application dérivée peut s'expliquer comme suit (*fig.* 49).

L'angle de deux courbes est défini comme l'angle de leurs tangentes; alors une transformation f est conforme si l'angle des courbes C et C' en z_0 peut être « transporté » sur l'angle des courbes images $f \circ C$ et $f \circ C'$ au point z_0.

Lorsque la fonction f est injective et holomorphe, on montre qu'elle définit un *homéomorphisme* de l'ouvert (supposé de plus connexe) D — où elle admet ces deux propriétés — sur l'ensemble f (D), qui est aussi un ouvert; de plus, l'application réciproque f^{-1} est holomorphe dans f (D). On dit, dans ce cas, que les domaines D et f (D) sont *conformément équivalents*, ou *isomorphes*. Très exactement, on dit qu'un homéomorphisme d'un ouvert D sur un ouvert D' défini par une fonction holomorphe est un *isomorphisme* de D sur D' lorsque l'application réciproque est aussi holomorphe (il s'agit, bien entendu, de prendre en considération non seulement les propriétés topologiques, mais aussi celles induites par l'analyse complexe).

Le problème fondamental de la représentation conforme est en quelque sorte la réciproque : deux ouverts D et D' du plan complexe, donnés, sont-ils isomorphes ?

Ses ramifications sont très vastes, et l'on peut citer en particulier la résolution de problèmes aux limites (problème de Dirichlet) relatifs à des potentiels (newtoniens par exemple).

Le problème n'est pas simple, car tout homéomorphisme de D sur D' n'est pas forcément un isomorphisme. Par exemple, le plan complexe $\mathbb{C}$ et le disque ouvert

$$D = \{z \in \mathbb{C} \text{ tels que } |z| < 1\}$$

sont homéomorphes mais pas isomorphes (car un isomorphisme de $\mathbb{C}$ sur D serait holomorphe et borné, donc constant, et par suite non injectif).

On montre d'abord que, si f est un isomorphisme de D sur D' et φ un isomorphisme de D sur D (ou *automorphisme* de D), alors $f \circ \varphi$ est encore un isomorphisme de D sur D'. Toutes les représentations conformes de D sur D' peuvent donc être obtenues à partir de l'une d'entre elles par composition avec un élément quelconque du groupe des automorphismes de D. La détermination de ce groupe peut donc être très utile. Ainsi, il est facile de voir que l'ensemble des automorphismes de C est formé des transformations $z \mapsto az + b$ où a est non nul; on peut aussi montrer que les automorphismes du disque unité ouvert D sont les transformations homographiques :

$$z \mapsto e^{i\theta}\left(\frac{z + z_0}{1 + \bar{z}_0 \cdot z}\right) \quad \text{où } \theta \in \mathbb{R} \text{ et } z_0 \in D \text{ sont arbitraires.}$$

Enfin le **théorème fondamental** (énoncé par Riemann) **de la représentation conforme** — tout ouvert $V \subset \mathbb{C}$ est isomorphe au disque ouvert unité D s'il est distinct de $\mathbb{C}$ et simplement connexe — permet de répondre au problème posé. Toutefois, il est à noter que la détermination approchée de la fonction holomorphe qui définit l'isomorphisme de V sur D est en général très difficile et relève du calcul numérique.

On peut dire que la représentation conforme est à la fois, parmi les problèmes relatifs aux fonctions de variable complexe, le plus ancien et l'un des plus pratiqués couramment. En effet, la projection stéréographique est une invention grecque et est l'un des outils privilégiés de la cartographie. Il s'agit (*fig. 50*) de l'application qui, à tout point M (x_1, x_2, x_3) de la sphère unité de $\mathbb{R}^3$ tel que $x_3 \neq 1$,

associe le point $M'\left(\dfrac{x_1}{1 - x_3}, \dfrac{x_2}{1 - x_3}\right)$ du plan $x_3 = 0$; ce

point M' n'est autre que l'intersection de la droite joignant le « pôle nord » de la sphère au point M avec le plan $x_3 = 0$. Si l'on identifie ce plan au plan complexe, on a l'application :

$$\sigma_1 : (x_1, x_2, x_3) \mapsto \frac{x_1}{1 - x_3} + i\frac{x_2}{1 - x_3}$$

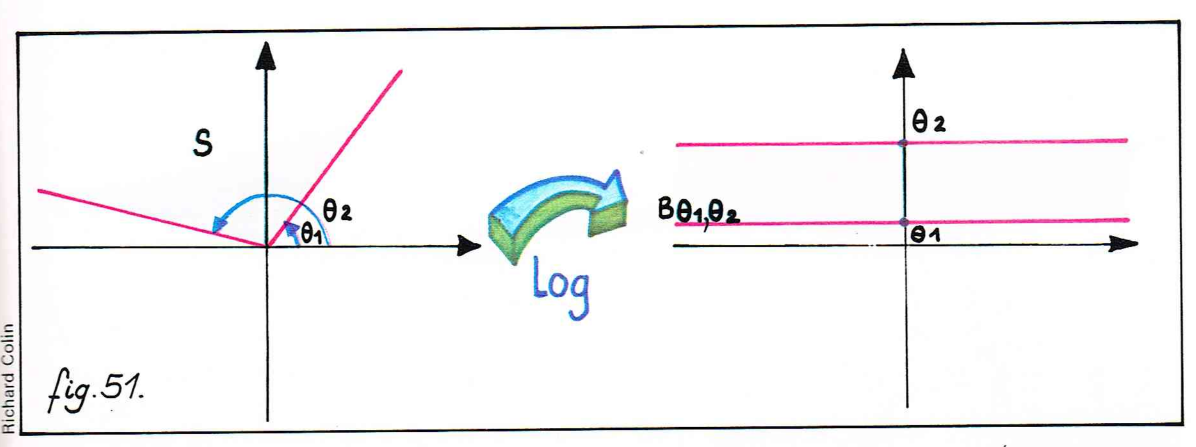
qui est une représentation conforme de la sphère privée du pôle $x_3 = 1$ sur le plan complexe.

On pourrait aussi bien définir la projection stéréographique relative au « pôle sud » $x_3 = -1$, qui, composée avec la symétrie par rapport à l'axe x_1, donne la représentation conforme :



▶ *Le mathématicien allemand* **Bernhard Riemann** *(1826-1866) : il développa l'idée d'une géométrie fondée sur l'hypothèse suivant laquelle par un point on ne peut mener aucune parallèle à une droite.*

Palais de la Découverte, Paris



◀ ***Figure 51 :*** ***représentation conforme du secteur S sur la bande B_{θ_1, θ_2} par la fonction logarithme complexe.***

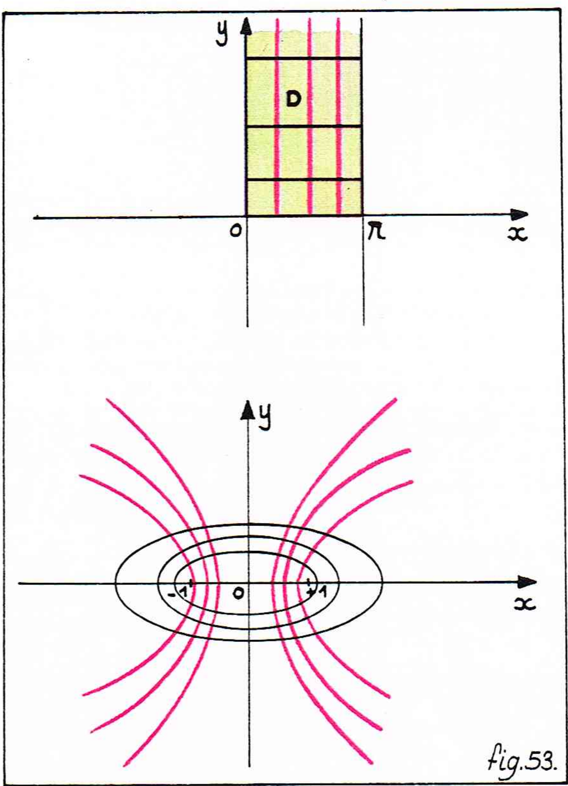
$$\sigma_2 : (x_1, x_2, x_3) \mapsto \frac{x_1}{1 + x_3} - i \frac{x_2}{1 + x_3}$$

de la sphère privée du pôle $x_3 = -1$ sur le plan C.

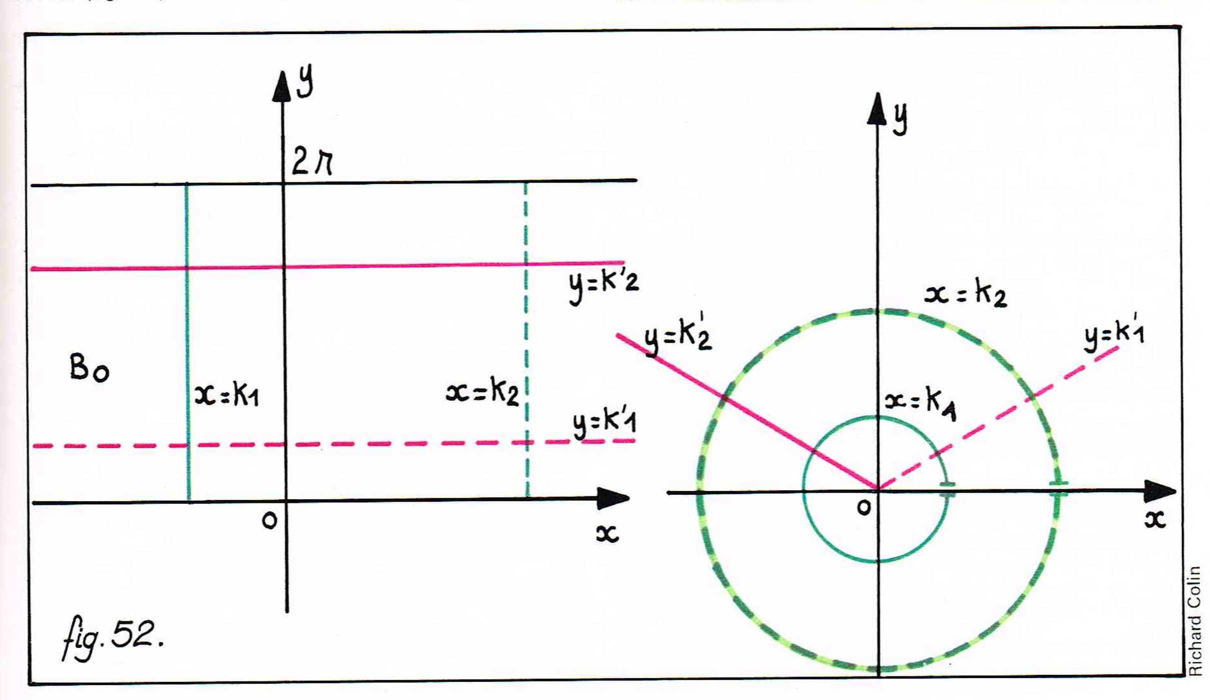
On peut encore donner quelques exemples de représentations conformes. La fonction $z \mapsto \log z$ est, on l'a vu, une représentation conforme du plan privé du demi-axe réel négatif sur la bande $B_{-\pi} = \{z \in C \text{ tels que } |Im\, z| < \pi\}$. C'est donc une représentation conforme (*fig.* 51) du « secteur » $S = \{z \in C \text{ tels que } \theta_1 < \text{Arg}\, z < \theta_2\}$ sur la bande $B_{\theta_1, \theta_2} = \{z \in \mathbb{C} \text{ tels que } \theta_1 < Im\, z < \theta_2\}$.

On peut donc considérer la transformation conforme réciproque réalisée par la fonction : $z \mapsto e^z$ dont la restriction à la bande : $B_0 = \{z \in \mathbb{C} \text{ tels que } 0 < Im\, z < 2\pi\}$ est injective, et qui envoie cette bande sur le plan privé (on dit aussi fendu le long) du demi-axe réel positif. Sur la figure 52, on a représenté la transformation sur les segments des droites $x = k$, inclus dans la bande et sur les droites $y = k'$ de la même bande, et dont les images respectives sont des cercles centrés à l'origine, privés de leur seul point réel positif d'une part et des demi-droites partant de l'origine.

De là, on peut par exemple étudier la transformation associée à $z \mapsto e^{iz}$. Puis, en composant avec celle définie par : $t \to \frac{1}{2}\left(t + \frac{1}{t}\right)$, on obtient la représentation conforme associée à : $z \mapsto \cos z$ sur l'ensemble ouvert $D = \{z \in \mathbb{C}$ tels que $0 < Re\, z < \pi$ et $Im\, z > 0\}$. En particulier, on peut voir que le segment de la droite $y = k$ inclus dans D est transformé en une ellipse dont les foyers sont les points réels $+1$ et -1 et que la demi-droite $x = k'$ incluse dans D est transformée en une hyperbole homofocale (*fig.* 53).



◀ ***Figure 53 :*** ***représentation conforme par la fonction cosinus. Les segments y = Cte de D deviennent des ellipses, et les demi-droites x = Cte de D des hyperboles homofocales.***



◀ ***Figure 52 :*** ***représentation conforme de la bande B_0 sur $\mathbb{C}$ privé de $\mathbb{R}_+$, par la fonction exponentielle; transformation des droites x = Cte et y = Cte.***

▶ *Figure 54; série de Fourier de la fonction 2 π-périodique égale à x sur]0, 2 π[:* $y = \pi - 2\left(\frac{\sin x}{1} + \frac{\sin 2x}{2} + \frac{\sin 3x}{3} + \ldots\right)$.

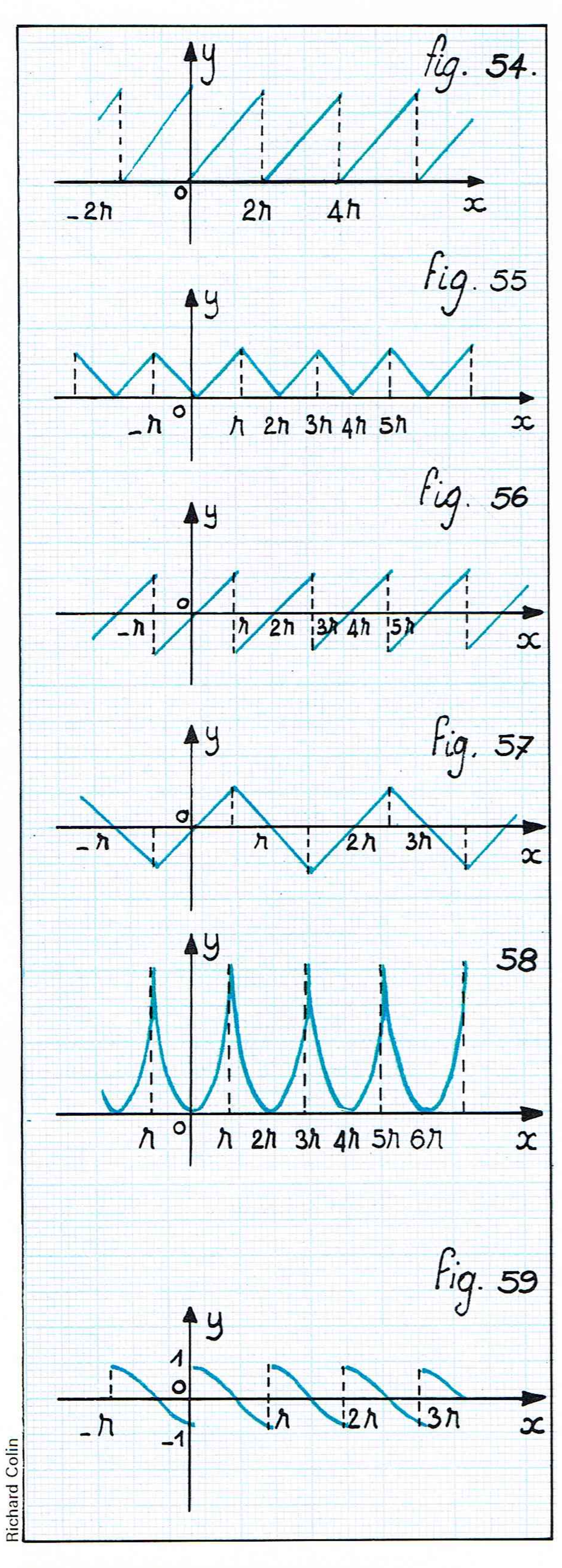
Figure 55; série de Fourier de la fonction 2 π-périodique égale à x sur [0, π] et à — x sur [— π, 0] : $y = \frac{\pi}{2} - \frac{4}{\pi}\left(\cos x + \frac{\cos 3x}{3^2} + \frac{\cos 5x}{5^2} + \ldots\right)$.

Figure 56; série de Fourier de la fonction 2 π-périodique égale à x sur]— π, + π[: $y = 2\left(\frac{\sin x}{1} - \frac{\sin 2x}{2} + \frac{\sin 3x}{3} - \ldots\right)$.

Figure 57; série de Fourier de la fonction 2 π-périodique égale à x sur $\left[-\frac{\pi}{2}, +\frac{\pi}{2}\right]$ *et à π — x sur* $\left[\frac{\pi}{2}, \frac{3\pi}{2}\right]$: $y = \frac{4}{\pi}\left(\sin x - \frac{\sin 3x}{3^2} + \frac{\sin 5x}{5^2} - \ldots\right)$.

Figure 58; série de Fourier de la fonction 2 π-périodique égale à x² sur [— π, + π] : $y = \frac{\pi^2}{3} - 4\left(\frac{\cos x}{1} - \frac{\cos 2x}{2^2} + \frac{\cos 3x}{3^2} - \ldots\right)$.

Figure 59; série de Fourier de la fonction 2 π-périodique égale à cos x sur]0, π[et à sin x sur]— π, 0[: $y = \frac{4}{\pi}\left(\frac{2 \sin 2x}{1 \cdot 3} + \frac{4 \sin 4x}{3 \cdot 5} + \frac{6 \sin 6x}{5 \cdot 7} + \ldots\right)$.



Analyse harmonique

Par les liens très étroits que l'on peut tracer entre l'analyse harmonique et la théorie de l'intégration, par ses applications en physique et par la contribution apportée au concept moderne de fonction, on peut dire que cette branche des mathématiques est la pierre de touche de l'analyse mathématique. C'est en étudiant l'équation des cordes vibrantes $\frac{\partial^2 y}{\partial t^2} = a^2 \frac{\partial^2 y}{\partial x^2}$ que le mathématicien suisse Daniel Bernoulli (1700-1782) introduisit une fonction (solution de cette équation) sous forme d'une série trigonométrique :

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left(a_n \cos \frac{n\pi}{l} x + b_n \sin \frac{n\pi}{l} x\right)$$

L'analogie acoustique a conduit à donner l'appellation « *harmonique* » aux termes de cette série autres que le « niveau fondamental » obtenu pour $n = 1$.

Le développement en série trigonométrique,

$$\sum_{n \geqslant 0} (a_n \cos nx + b_n \sin nx) \quad \text{ou} \quad \sum_{n \geqslant 0} \rho_n \cos (nx + \varphi_n),$$

ou encore $\sum_{n \in \mathbb{Z}} c_n e^{inx}$, des fonctions périodiques allait alors permettre de considérer les fonctions indépendamment des propriétés de continuité, de dérivabilité, etc. C'est ainsi que J. Fourier essaya de représenter des fonctions discontinues comme somme de séries trigonométriques (*fig.* 54, 55, 56, 57, 58, 59) ; le même problème que celui posé par la série de Taylor, c'est-à-dire la convergence de la série de Fourier d'une fonction vers cette fonction, ouvrait alors tous les développements de l'analyse harmonique et de la théorie spectrale.

Séries trigonométriques

Soit f une fonction de variable réelle, 2 π-périodique (cette valeur n'étant évidemment pas restrictive pour ce qui suit), que l'on suppose intégrable. On appelle *coefficients de Fourier* de f les nombres $c_n(f) = \frac{1}{2\pi}\int_0^{2\pi} f(x)\, e^{-inx}\, dx$ (si la fonction était de période T, on poserait

$$c_n(f) = \frac{1}{T}\int_0^{T} f(x)\, e^{-\frac{2\pi}{T}inx}\, dx)$$

définis pour tout $n \in \mathbb{Z}$, et on dira que la série

$$\sum_{n \in \mathbb{Z}} c_n(f)\, e^{inx}$$

est la *série de Fourier* de f. Le problème étant de savoir dans quel cas cette série est convergente de somme f (ponctuellement ou uniformément selon les cas), on est amené à étudier l'influence des propriétés éventuelles de f (continuité, différentiabilité, etc.) sur la convergence de la série et à introduire un nouveau type de convergence par les moyennes arithmétiques des sommes de Fourier que l'on appelle les *sommes de Féjer :*

$$\sigma_N(f, x) = \frac{S_0(f, x) + S_1(f, x) + \ldots + S_{N-1}(f, x)}{N} = \sum_{k=-N}^{+N}\left(1 - \frac{|k|}{N}\right) c_k(f)\, e^{ikx}$$

où l'on pose $S_n(f, x) = \sum_{k=-n}^{+n} c_k(f)\, e^{ikx}$ *(somme de Fourier)*.

Les hypothèses fortes que l'on peut (dans certains cas) faire pour la fonction f donnent les résultats suivants :

— si la fonction f est deux fois continûment dérivable, alors la série de Fourier de f converge absolument et uniformément vers f ;

— si la fonction f est différentiable, sa série de Fourier converge uniformément vers f ;

— si la série de Fourier de f est normalement convergente (c'est-à-dire si $\sum_{n \in \mathbb{Z}} |c_n(f)|$ converge), f est somme de sa série de Fourier.

En affaiblissant les hypothèses, on obtient alors que, si f est une fonction continue, les sommes de Fourier $S_n(f, x)$ ne convergent vers $f(x)$ que presque partout (c'est-à-dire à l'exception de points formant un ensemble dont la mesure est nulle) ; par contre, les *sommes de Féjer* $\sigma_N(f, x)$ convergent uniformément vers $f(x)$. Il existe des fonctions continues qui ne sont pas égales à la somme de leur série de Fourier, comme l'a montré Féjer avec la fonction $f(t) = \sum_{k=1}^{\infty} \frac{1}{k^2} e^{2iN_k t} \cdot F_{N_k}(t)$ où, pour

tout k, N_k est un entier vérifiant $\lim_{k \to \infty} N_k = +\infty$ (suite d'entiers croissant indéfiniment) et où F_{N_k} est le « polynôme trigonométrique » $F_{N_k}(t) = \sum_{n=1}^{N_k} \frac{\sin nt}{n}$; pour cette fonction f, on a $c_n(f) = \sum_{k=1}^{\infty} \frac{1}{k^2} c_{n-2N_k}(F_{N_k})$, et la série de Fourier ainsi définie diverge au point 0.

Les fonctions de carré intégrable (voir le paragraphe *Intégration*) possèdent une propriété de convergence pour leur série de Fourier, mais relative à la structure hilbertienne de l'espace L^2, c'est-à-dire que, si f est une fonction 2π-périodique, $f \in L^2[0, 2\pi]$, on a :

$$\lim_{N \to \infty} \left[\int_0^{2\pi} |f(x) - \sum_{-N}^{+N} c_n(f) e^{inx}|^2 dx \right] = 0$$

Pour les fonctions de cette classe, on possède encore les remarquables *formules de Parseval :*

$$\frac{1}{2\pi} \int_0^{2\pi} f(t) g(t) dt = \sum_{n \in \mathbb{Z}} c_n(f) c_{-n}(g)$$

$$\frac{1}{2\pi} \int_0^{2\pi} |f(t)|^2 dt = \sum_{n \in \mathbb{Z}} |c_n(f)|^2$$

vérifiées pour tout $f \in L^2[0, 2\pi]$ et tout $g \in L^2[0, 2\pi]$.

La transformation de Fourier

Le développement d'une fonction en série de Fourier fait apparaître une correspondance qui, à une fonction f, 2π-périodique et intégrable, associe une fonction $c_n(f)$ définie pour $n \in \mathbb{Z}$. On va généraliser cette transformation, associant à toute fonction f intégrable son *image* (ou *transformée*) *de Fourier* F définie par :

$$F(t) = \int_{-\infty}^{+\infty} e^{-2\pi itx} f(x) dx;$$

on note souvent $F = \mathcal{F}f$. Ceci permet d'obtenir un outil d'analyse applicable à des fonctions pour lesquelles les méthodes des séries de Fourier sont inadaptées (éléments de l'espace $L^1[0, 2\pi]$ par exemple). Il faut noter que le coefficient -2π est placé dans cette définition, bien que son usage ne soit pas totalement généralisé ; son plus grand intérêt réside dans la simplicité de la propriété de « réciprocité » qu'on expose plus avant.

Si f est une fonction intégrable, sa transformée de Fourier est continue, bornée et tend vers zéro à l'infini : $\lim_{|t| \to \infty} F(t) = 0$. Cette « régularité » est d'autant plus accentuée que la fonction f est elle-même régulière ; ainsi, lorsque f est p fois continûment dérivable, on a $\mathcal{F}f^{(p)} = (2\pi it)^p \mathcal{F}f$ et, de plus

$$|F(t)| \leq \frac{1}{|2\pi t|^p} \int_{-\infty}^{+\infty} |f^{(p)}(x)| dx,$$

ce qui montre qu'alors $\mathcal{F}f = F$ tend vers zéro à l'infini de façon encore plus marquée.

Réciproquement, si $x^p f(x)$ est intégrable, alors $\mathcal{F}f$ est p fois continûment dérivable, et l'on a :

$$(\mathcal{F}f)^{(p)}(t) = \mathcal{F}[(-2\pi ix)^p f(x)].$$

Enfin, une étude plus précise montre que la transformée de Fourier d'une fonction se « disperse » d'autant plus que la fonction est « tassée » (en ce qui concerne leurs valeurs).

On introduit, parallèlement, la *transformation de Fourier conjuguée*, $\overline{\mathcal{F}}$, souvent appelée *réciproque*, définie pour toute fonction intégrable par

$$\overline{\mathcal{F}}f(t) = \int_{-\infty}^{+\infty} e^{2\pi itx} f(x) dx.$$

On peut alors dégager l'aspect essentiel de la transformation de Fourier ; en effet, si une fonction f est intégrable et si sa transformée $\mathcal{F}f$ l'est aussi, on peut écrire que $f = \overline{\mathcal{F}}(\mathcal{F}f)$. En d'autres termes, la transformation de Fourier — pour cette catégorie de fonctions — caractérise une fonction à partir de sa transformée, et *vice versa :*


Palais de la Découverte, Paris

◀ ***Le mathématicien français* Joseph Fourier *(1768-1830) ; ses travaux conduisirent à l'une des grandes études mathématiques du XIXe siècle : celle des séries dites de Fourier.***

c'est la *propriété de réciprocité.* Dans le cas des distributions (voir *Analyse fonctionnelle*), on parlera pour ces transformations $\mathcal{F}$ et $\overline{\mathcal{F}}$ d'isomorphismes continus de l'espace des distributions tempérées.

La propriété fondamentale de la transformation de Fourier est liée à la notion de produit de convolution. On peut considérer les coefficients de la série de Fourier d'une fonction comme une transformée de Fourier définie sur $\mathbb{Z}$ en posant $\hat{f}(k) = c_k(f)$; or la première formule de Parseval peut s'écrire :

$$c_0(fg) = \sum_{n \in \mathbb{Z}} c_n(f) c_{-n}(g) \quad \text{et} \quad c_m(fg) = c_0(fge^{imt})$$

implique $c_m(fg) = \sum_{n \in \mathbb{Z}} c_n(f) c_{m-n}(g)$, ce que nous écrirons donc $\widehat{fg}(m) = \sum_{n \in \mathbb{Z}} \hat{f}(n) \hat{g}(m-n)$. D'autre part, cette même formule de Parseval peut s'écrire :

$$\sum_{n \in \mathbb{Z}} c_n(f) c_n(g) e^{int} = \frac{1}{2\pi} \int_0^{2\pi} f(x) g(t-x) dx$$

On voit donc apparaître une loi de composition entre fonctions définies sur un ensemble discret dans le premier cas, sur un ensemble continu dans le second ; on abordera seulement ce dernier cas. Soit f et g deux fonctions intégrables sur $\mathbb{R}$, on appelle *produit de convolution* de f et g la fonction $f * g$ définie par :

$$f * g(t) = \int_{-\infty}^{+\infty} f(x) g(t-x) dx$$

et l'on montre qu'elle est intégrable. Dans ces conditions, on peut énoncer que $\mathcal{F}(f * g) = (\mathcal{F}f)(\mathcal{F}g)$ et $\mathcal{F}(fg) = (\mathcal{F}f) * (\mathcal{F}g)$. La première de ces formules n'est autre que la formule générale de celle trouvée plus haut pour les coefficients de la série de Fourier, soit $\widehat{fg}(m) = \sum_{n \in \mathbb{Z}} \hat{f}(n) \hat{g}(m-n)$. Cette propriété trouve des applications très nombreuses tant dans la résolution des équations aux dérivées partielles qu'en calcul des probabilités.

La transformation de Fourier trouve une cohérence particulièrement nette avec la théorie de l'intégration et le problème de la mesure de Haar dans le cadre des groupes abéliens localement compacts (par exemple $\mathbb{R}$ ou $\mathbb{R}^m$, ou encore le tore $\mathbb{T} = \frac{\mathbb{R}}{2\pi\mathbb{Z}}$ ou encore $\mathbb{Z}$) ; son extension aux fonctions généralisées ou distributions permet de précieuses utilisations en physique.



Giraudon

▲ ***Portrait du mathématicien et physicien français* Pierre Simon de Laplace *(1749-1827).***

La transformation de Laplace

La transformation de Fourier peut être étendue aux fonctions de plusieurs variables et donne les mêmes résultats principaux. Elle peut être encore étendue au plan complexe, mais en ne considérant cette fois que des fonctions d'une variable réelle, nulles en dehors de l'ensemble $[0, +\infty[$. En raison de cette restriction, on a recours à la *fonction* dite *d'Heaviside*, notée Y, telle que $Y(t) = 0$ si $t < 0$, et $Y(t) = 1$ si $t \geqslant 0$. Si f est donc une fonction nulle hors de $[0, +\infty[$ — si ce n'est pas le cas, on considère alors $Y(t)\,f(t)$ — on appelle *transformée de Laplace de f* la fonction de la variable complexe z définie par : $\mathfrak{L}f(z) = \int_0^{+\infty} f(t)\,e^{-zt}\,dt$. Ainsi, la fonction d'Heaviside Y a pour transformée $\frac{1}{z}$, ce que l'on note $\frac{1}{z} \sqsubset Y(t)$ au lieu de $\frac{1}{z} = \mathfrak{L}Y$; de même, on voit aisément que $e^{-az} \sqsubset \delta$ si $\delta(t) = 1$ pour $t = a \geqslant 0$ et $\delta(t) = 0$ sinon. Un calcul simple montre que, si $z = x + iy$ en fixant x,

$$\mathfrak{L}f(z) = \mathcal{F}\,[f(t)\,e^{-tx}]\left(\frac{y}{2\pi}\right)$$

Le domaine de définition de $\mathfrak{L}f$ est déterminé par le fait que, si $\int_0^{+\infty} f(t)\,e^{-tz_0}\,dt$ converge absolument, alors $\int_0^{+\infty} f(t)\,e^{-tz}\,dt$ converge absolument et uniformément dans le demi-plan fermé $Re(z) \geqslant Re(z_0)$ et dans l'angle $|Arg(z - z_0)| \leqslant \theta$ pour $\theta \in \left[0, \frac{\pi}{2}\right[$ si $\int_0^{+\infty} f(t)\,e^{-tz_0}\,dt$ converge simplement. On définit alors l'*abscisse de convergence absolue* comme l'élément unique $\alpha \in \overline{\mathbb{R}}$ tel que :

$$Re(z) > \alpha \Rightarrow \int_0^{+\infty} |f(t)\,e^{-tz}|\,dt < +\infty$$

et $$Re(z) < \alpha \Rightarrow \int_0^{+\infty} |f(t)\,e^{-tz}|\,dt = +\infty$$

et l'*abscisse de convergence* comme l'élément unique $\beta \in \overline{\mathbb{R}}$ tel que :

$$Re(z) > \beta \Rightarrow \int_0^{+\infty} f(t)\,e^{-tz}\,dt \text{ converge}$$

et $$Re(z) < \beta \Rightarrow \int_0^{+\infty} f(t)\,e^{-tz}\,dt \text{ diverge};$$

il est clair que $\beta \leqslant \alpha$. Par exemple, l'abscisse de convergence absolue de $Y(t)$ est $\alpha = 0$.

On montre aussi que la transformée de Laplace est holomorphe dans le demi-plan de convergence,

$$Re(z) > \beta;$$

dans ce demi-plan, si $F(z) \sqsubset f(t)$, on a $F^{(k)}(z) \sqsubset (-t)^k f(t)$ pour tout k.

Enfin, comme pour la transformation de Fourier, on a la propriété fondamentale :

$$\mathfrak{L}(f * g) = (\mathfrak{L}f) \cdot (\mathfrak{L}g).$$

C'est cette propriété qui donne lieu au plus important développement de la transformation de Laplace, le *calcul symbolique*, destiné à résoudre, après lecture des tables de transformées, les équations de convolution de la forme $f * g = h$ où g est inconnue (par transformation de Laplace, ces équations s'écrivent $\mathfrak{L}f \cdot \mathfrak{L}g = \mathfrak{L}h$, soit encore $\mathfrak{L}g = \frac{\mathfrak{L}h}{\mathfrak{L}f}$). Ainsi, pour déterminer f telle que $\int_0^t f(x)\,e^{k(t-x)}\,dx = t$, on pose l'équation sous la forme : $Y(x)\,e^{kx} * f(x) = tY(t)$; en prenant les transformées de Laplace, on obtient $\frac{1}{z-k} \cdot \mathfrak{L}f = \frac{1}{z^2}$, soit $\mathfrak{L}f = \frac{1}{z} - \frac{k}{z^2}$; par conséquent, la transformation de Laplace étant injective, on a $f(x) = Y(x)(1 - kx)$.

Ce calcul symbolique est d'un usage très courant (particulièrement dans les problèmes d'automatique, où les problèmes sont résolus à l'aide d'une transformation opérant sur une variable discrète, la *transformation en z* qui a de multiples applications, économiques par exemple, dès que l'on aborde des modèles dits « avec retard »), en raison de sa simplicité (on se borne en général à lire des tables) et du formalisme très simple qui en est la base et qui trouve une très belle application dans le cas des mesures et des distributions.

BIBLIOGRAPHIE

BOREL E., *Leçons sur la théorie des fonctions,* Gauthier-Villars, Paris. - BOURBAKI N., *Éléments d'histoire des mathématiques,* Hermann ; *Fonctions d'une variable réelle,* Hermann ; *Intégration,* Hermann. - CARNOT L., *la Métaphysique du calcul infinitésimal,* Gauthier-Villars. - CARTAN H., *Calcul différentiel,* Hermann ; *Théorie élémentaire des fonctions analytiques d'une ou plusieurs variables complexes.* Hermann. - DELACHET A., *l'Analyse mathématique,* P. U. F., « Que sais-je ? » ; *Calcul différentiel et intégral,* P. U. F., « Que sais-je ? ». - DIEUDONNÉ J., *Calcul infinitésimal,* Hermann. - FÉLIX L., *Exposé moderne des mathématiques élémentaires,* Dunod. - LE LIONNAIS F. (sous la direction de), *les Grands Courants de la pensée mathématique,* Blanchard. - LELONG-FERRAND J. et ARNAUDIÈS J.-M., *Cours de mathématiques,* t. II, Dunod. - RIESZ F. et NAGY B., *Leçons d'analyse fonctionnelle.* — SAKS B. et ZYGMUND A., *Analytic Functions.* - SCHWARTZ L., *Cours d'analyse,* t. I et II, Hermann ; *Méthodes mathématiques pour les sciences physiques,* Hermann.

ANALYSE FONCTIONNELLE

Espaces fonctionnels

La notion de « fonction » est fondamentale en mathématiques. Ainsi qu'on l'a vu dans le chapitre *Langage ensembliste* une fonction (ou plus généralement une application) est définie comme un procédé « qui met en correspondance » un ensemble E et un ensemble F. Ce procédé associe à tout élément x de E un *élément unique* y de F dont la dépendance à l'égard de x est symbolisée par l'écriture $y = f(x)$. La lettre f représente le procédé en question. L'on parlera donc de la fonction $f : E \to F$.

Il s'agit donc d'une notion suffisamment générale pour être utilisée selon des points de vue très divers. Les algébristes s'intéressent plutôt aux *propriétés globales* des fonctions; ils les définissent et les étudient dans la mesure où celles-ci « transportent » une structure algébrique d'un ensemble vers un autre (homomorphismes, isomorphismes, etc.). Par contre les analystes s'intéressent plus volontiers aux *propriétés locales*. Ainsi étudient-ils les points de singularité des fonctions, c'est-à-dire les points de l'ensemble de départ où « il se passe quelque chose » (discontinuités, comportement asymptotique, etc.). Mais algébristes et analystes s'accordent sur un point : une fois qu'ils ont jeté leur dévolu sur une fonction f, ils ne sont guère enclins à élargir leur champ d'étude en considérant d'autres fonctions qui, au regard de certains critères, « ressemblent » à f.

A l'inverse le point de vue de l'analyse fonctionnelle consiste à remarquer que l'écriture $f(x)$ présente une symétrie entre « f » et « x ». Dès lors il est tentant de considérer également « f » comme variable.

L'analyse fonctionnelle est donc l'étude des ensembles de fonctions (ou espaces fonctionnels) et des structures algébriques et surtout topologiques dont on peut les munir. Nous verrons d'abord comment un espace fonctionnel est, le plus souvent, un *espace vectoriel topologique* (en abrégé *e. v. t.*). Procédant du général vers le particulier, nous verrons apparaître ensuite : les espaces localement convexes, les espaces de Banach, les espaces de Banach réflexifs, les espaces de Hilbert, enfin les espaces de dimension finie, c'est-à-dire $\mathbb{R}^n$. Ainsi donc, l'analyse fonctionnelle développe des concepts permettant de rassembler dans un même formalisme la topologie, l'analyse, d'une part, et la géométrie et l'algèbre, d'autre part. A cet égard, l'analyse fonctionnelle est d'un intérêt théorique tout à fait évident. Mais elle répond aussi aux besoins de l'analyse numérique : une équation aux dérivées partielles, par exemple, est une équation dans laquelle l' « inconnue » est une *fonction* (d'une ou de plusieurs variables), et non plus un nombre ou un vecteur comme dans les équations algébriques; il importe donc de déterminer *a priori* l'espace fonctionnel auquel appartiendront la ou les solutions de l'équation. D'autre part, la résolution pratique des problèmes d'analyse numérique nécessite la mise en œuvre de *procédés d'approximation :* « approcher » une fonction suppose que l'on soit en mesure de définir l'idée de « proximité » — c'est-à-dire la notion de voisinage — dans un espace fonctionnel. C'est dire l'intérêt qu'il peut y avoir à étudier les espaces fonctionnels d'un point de vue topologique.

Soit donc E et F deux ensembles; nous noterons F^E l'ensemble de toutes les applications de E dans F. Cet ensemble est particulièrement vaste et, suivant la nature du problème à résoudre, nous nous restreindrons à tel ou tel sous-ensemble de F^E; il convient d'ores et déjà de faire la remarque suivante : si l'ensemble d'arrivée F est muni d'une certaine structure (algébrique ou topologique), il est souvent possible de munir F^E d'une structure analogue. L'ensemble de départ E joue un rôle secondaire dans la définition de structures sur l'ensemble F^E : ainsi, si F est un espace vectoriel (sur $\mathbb{R}$), F^E est aussi un espace vectoriel pour les deux opérations :

$$\begin{aligned} (f+g)(x) &= f(x) + g(x) \qquad \forall x \in E \\ \lambda \cdot f(x) &= \lambda \cdot f(x) \qquad \forall x \in E, \forall \lambda \in \mathbb{R} \end{aligned}$$

Supposons maintenant que F est un espace métrique; il semble légitime d'espérer que F^E puisse être lui aussi un espace métrique. Cela n'est pas le cas, car il n'est pas possible de munir F^E d'une *distance*. Cependant, en se restreignant au sous-ensemble $(F^E)_b$ des *applications bornées*, on peut obtenir ce résultat : soit en effet d la distance définie sur F, et f et g deux fonctions bornées de F^E; il est aisé de démontrer que la quantité

$$\delta(f, g) = \sup_{x \in E} d(f(x), g(x))$$

est définie et qu'elle permet de définir une distance sur $(F^E)_b$.

Il est facile de déduire de ces deux résultats que, lorsque F est un espace vectoriel normé (voir *Topologie*), alors l'ensemble $(F^E)_b$ est aussi un espace vectoriel normé en posant :

$$|||f||| = \sup_{x \in E} ||f(x)||$$

Les *espaces métriques complets* sont très appréciés en analyse fonctionnelle : rappelons que, dans de tels espaces, on est assuré de la convergence d'une suite, même si on est dans l'impossibilité d'en déterminer la limite, pourvu qu'il s'agisse d'une suite de Cauchy. Des résultats importants d'analyse sont établis dans le cadre des espaces complets. Il est donc intéressant de savoir dans quelle mesure un espace fonctionnel peut être un espace complet.

On démontre le résultat suivant : *si F est un espace métrique complet, l'espace métrique $(F^E)_b$ est aussi complet.*

Exemple : considérons l'ensemble des fonctions continues de $[a, b] \subset \mathbb{R} \to \mathbb{R}$. Il s'agit de fonctions bornées (l'intervalle $[a, b]$ étant compact, les fonctions continues y sont bornées). Nous avons donc affaire à une espace normé, complet, c'est-à-dire à un *espace de Banach*.

Convergence simple et uniforme d'une suite de fonctions

Nous allons définir deux types de convergence de fonctions : la *convergence simple* et la *convergence uniforme*, et nous montrerons que ces deux types de convergence correspondent à deux topologies différentes de l'espace (F^E) ou $(F^E)_b$.

L'ensemble F sera dorénavant supposé être un espace métrique.

Convergence simple

Nous dirons qu'une suite de fonctions f_n *converge simplement* vers une fonction f si, pour tout $x \in E$,

$$f_n(x) \to f(x)$$

ce qui s'écrit : **(1)**

$$\forall x \in E \quad \forall \varepsilon > 0 \quad \exists m \in \mathbb{N} / \forall n \geqslant m : d[f_n(x), f(x)] \leqslant \varepsilon.$$

L'on notera que l'entier m ainsi déterminé dépend non seulement de ε mais aussi de x. Cette notion de convergence correspond à la *topologie produit* de l'espace F^E. Cette topologie n'est pas métrisable, car elle ne peut pas être définie à l'aide d'une distance (sauf si l'on se restreignait aux fonctions bornées) ; par contre, nous verrons plus loin qu'il est possible de la définir à l'aide d'une *famille de semi-distances.*

Convergence uniforme

Nous dirons que la suite de fonctions f_n *converge uniformément* vers la fonction f si l'entier m déterminé dans l'expression **(1)** peut être choisi indépendamment de x.

Ce qui s'écrit :

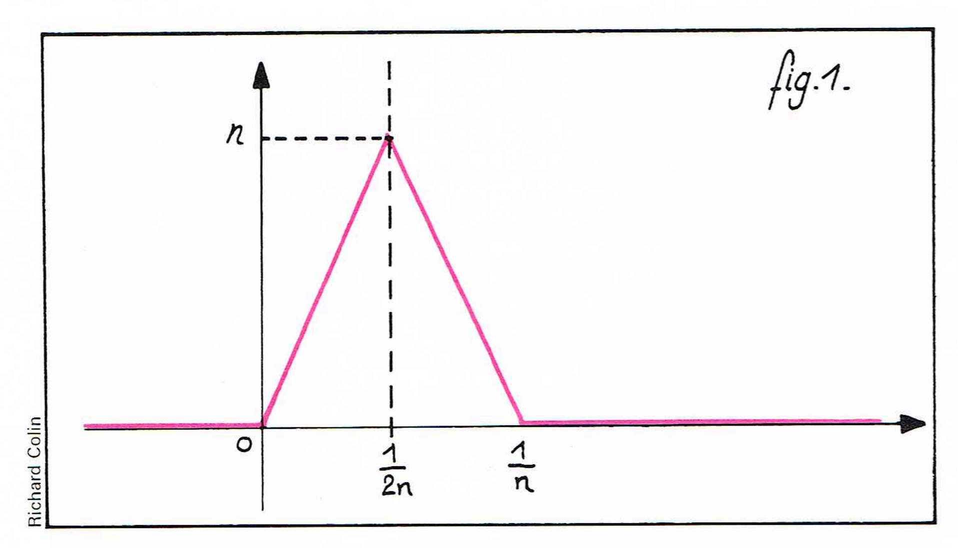
$$\forall x \in E \quad \forall \varepsilon > 0 \quad \exists m \in \mathbb{N} \quad \forall n \geqslant m : d(f_n(x), f(x)) \leqslant \varepsilon$$

Si les f_n et f sont des fonctions bornées, cette expression peut s'écrire :

$$\forall \varepsilon > 0 \quad \exists m \in \mathbb{N} \quad \forall n \geqslant m : \delta(f_n, f) \leqslant \varepsilon.$$

Dans ce cas, la convergence uniforme n'est autre que la convergence des points f_n de l'espace métrique $(F^E)_b$ vers le point f.

La convergence uniforme entraîne, bien sûr, la convergence simple. Mais la réciproque est fausse, ainsi qu'en témoigne le contre-exemple suivant : soit la suite de fonctions f_n de $\mathbb{R} \to \mathbb{R}$ définie par :



▲ *Figure 1 : graphe de la fonction énoncée dans le texte ci-contre.*

$$f_n(x) = \begin{cases} 0 & x \leqslant 0 \\ 2n^2 \cdot x & 0 \leqslant x \leqslant \frac{1}{2n} \\ -2n^2\left(x - \frac{1}{n}\right) & \frac{1}{2n} \leqslant x \leqslant \frac{1}{n} \\ 0 & x \geqslant \frac{1}{n} \end{cases}$$

Le graphe de cette fonction est donné par la *figure* 1. On montrera que la suite de f_n converge simplement vers la fonction 0. Par contre, elle ne converge pas uniformément vers 0 puisque

$$\delta(f_n, 0) = \operatorname*{Sup}_{x \in \mathbb{R}} |f_n(x) - 0| = n \to +\infty.$$

La principale vertu de la convergence uniforme est qu'elle « conserve » la notion de continuité par passage à la limite ; en d'autres termes, on démontre que : *la limite uniforme d'une suite de fonctions continues est continue.* L'énoncé d'un tel résultat suppose que l'espace de départ E soit un espace topologique (de façon à pouvoir parler de fonctions continues de E → F). La convergence uniforme est essentielle dans la démonstration de ce résultat : considérons par exemple la suite des fonctions φ_n : $[0, 1], \to \mathbb{R}$ définie par $\varphi_n(x) = x^n$; φ_n est continue pour tout n et converge *simplement* vers la fonction φ discontinue définie par :

$$\begin{cases} \varphi(x) = 0 & 0 \leqslant x < 1 \\ \varphi(1) = 1. \end{cases}$$

D'un point de vue topologique, ce résultat exprime le fait que dans l'espace $(\mathrm{FE})_b$, le sous-espace $(\mathrm{FE})_{bc}$ des applications bornées et continues est *fermé pour la topologie de la convergence uniforme.*

Espaces vectoriels topologiques

Définition

Un espace fonctionnel est le plus souvent un espace vectoriel ; ainsi que nous venons de le voir, il est possible de le munir d'une ou de plusieurs topologies. L'on choisira la topologie la mieux adaptée au problème que l'on veut résoudre, sachant qu'il ne sera pas toujours possible d'en trouver une qui soit métrisable (voir *Topologie de la convergence simple*). Ainsi donc, le cadre conceptuel de l'analyse fonctionnelle se révèle être naturellement celui des *espaces vectoriels topologiques.* Étant donné un espace vectoriel, il s'agit de déterminer quelles relations de compatibilité doivent exister entre sa structure algébrique et la structure topologique dont on souhaite le munir. Nous verrons qu'un espace vectoriel topologique n'est pas un espace vectoriel muni de n'importe quelle topologie.

Les espaces vectoriels sont en général de dimension infinie. C'est dire l'importance qu'il faut attacher au **théorème de F. Riesz** qui donne une caractérisation topologique des e. v. t. de dimension finie. Soit donc un espace vectoriel E muni d'une topologie $\mathcal{T}$ (métrisable ou non). On dira que E est un espace vectoriel topologique si

l'addition $\qquad (\vec{x}, \vec{y}) \in E \times E \longrightarrow \vec{x} + \vec{y}$

et la multiplication $\quad (\lambda, \vec{x}) \in \mathbb{R} \times E \longrightarrow \lambda \cdot \vec{x}$

sont des *applications continues* à *valeur dans E.* On dira alors que la structure vectorielle est compatible avec la structure topologique.

Nous avons rencontré en topologie des espaces vectoriels normés ; il est réconfortant de constater qu'un espace vectoriel normé est un espace vectoriel topologique. Ceci tient aux deux inégalités :

$$\begin{cases} \|(\vec{x} + \vec{y}) - (\vec{a} + \vec{b})\| \leqslant \|\vec{x} - \vec{a}\| + \|\vec{y} - \vec{b}\| & \vec{x}, \vec{y} \in E \\ \|\lambda\vec{x} - \alpha\vec{a}\| \leqslant |\lambda - \alpha| + |\alpha| \ \|\vec{x} - \vec{a}\| & \lambda, \alpha \in \mathbb{R}. \end{cases}$$

La première permet d'établir la continuité de l'addition, la seconde celle de la multiplication.

Propriétés des voisinages de $\vec{O}$ dans un espace vectoriel topologique

Voyons comment l'ensemble des voisinages de l'origine $\vec{O}$ suffit à caractériser la topologie de l'espace. Cela tient au fait que la structure vectorielle induit une sorte de « solidarité » entre tous les points de l'espace et qu'ainsi « ce qui se passe » à l'origine préfigure « ce qui se passe ailleurs ». Définissons préalablement ce que l'on appelle une partie *équilibrée et absorbante* de E (*fig.* 2).

Une partie A d'un espace vectoriel est dite *équilibrée* si, quels que soient $\lambda \in \mathbb{R}$ avec $|\lambda| \leqslant 1$ et $\vec{x} \in A$, on a $\lambda\vec{x} \in A$; géométriquement, on peut dire qu'une partie est équilibrée si elle contient la totalité du segment joignant tout point $\vec{x}$ à son symétrique $-\vec{x}$.

A sera dite *absorbante* si, pour tout $\vec{x}$, il existe un nombre $\alpha > 0$ tel que $\forall\lambda : |\lambda| \leqslant \alpha$ entraîne $\lambda\vec{x} \in A$; c'est-à-dire qu'il est alors toujours possible de « faire rentrer » n'importe quel point de E dans A par une homothétie de rapport suffisamment petit.

Nous allons montrer que la topologie de E est entièrement caractérisée par la donnée d'une famille de voisinages de $\vec{O}$ équilibrés et absorbants. Ce résultat s'énonce sous la forme d'un théorème :

Théorème : *soit* $\mathcal{F}(\vec{a})$ *l'ensemble de tous les voisinages d'un point* $\vec{a}$ *de* E ; *alors :*

(2) $\mathcal{F}(\vec{a}) = \vec{a} + \mathcal{F}(\vec{O})$;

(3) *tout voisinage* $V \in \mathcal{F}(\vec{O})$ *est absorbant ;*

(4) *il existe un système fondamental de voisinage de* $\vec{O}$ *équilibrés.*

Le **(2)** signifie simplement que l'on obtient n'importe quel voisinage V_a de $\vec{a}$ en faisant subir la translation $\vec{a}$ à un voisinage V_o de l'origine.

Pour démontrer **(3)** il suffit d'écrire que la multiplication par un scalaire est une opération continue de $\mathbb{R} \times E \to E$. Démontrons le dernier *item* plus en détail : soit V un voisinage de $\vec{O}$; montrons que l'on peut trouver un voisinage W de O, *équilibré*, inclus dans V (*cf. Topologie*). L'application $(\lambda, \vec{x}) \to \lambda \cdot \vec{x}$ est continue à l'origine : quel que soit donc le voisinage V de O, il existe $\alpha > 0$ et un voisinage W_1 de O tel que λW_1 soit inclus dans V pour tout λ tel que $|\lambda| < \alpha$. Posons $W = \bigcup_{|\lambda| < \alpha} \lambda W_1$; W est bien sûr un voisinage de O inclus dans V ; il nous reste à montrer qu'il est équilibré : soit $\vec{z} \in W$; montrons que $\forall\mu$ tel que $|\mu| < 1$, $\mu\vec{z} \in W$. D'après la définition de W, il existe λ ($|\lambda| \leqslant \alpha$) et $\vec{x} \in W_1$ tels que

$$z = \lambda \cdot \vec{x}; \quad \mu\vec{z} = \lambda \cdot \mu\vec{x}.$$

Or $|\lambda| < \alpha$ et $|\mu| \leqslant 1$ entraînent que $|\lambda \cdot \mu| < \alpha$ et donc que $\mu\vec{z} \in W$. W est donc un voisinage équilibré de $\vec{O}$.

De ce résultat fondamental, on peut tirer un certain nombre de conclusions : en particulier, on montrera que si une *application* linéaire d'un e. v. t. E dans un autre F

fig. 2.

ensemble équilibré et absorbant — ensemble non équilibré et absorbant — ensemble non absorbant et équilibré — ensemble non absorbant et non équilibré

◄ ***Figure 2 :** représentation graphique d'une partie* **équilibrée et absorbante** *de E.*

est continue à l'origine, elle est continue partout. Cette proposition est spécialement intéressante pour des espaces de dimension infinie (en dimension finie, toute application linéaire est nécessairement continue).

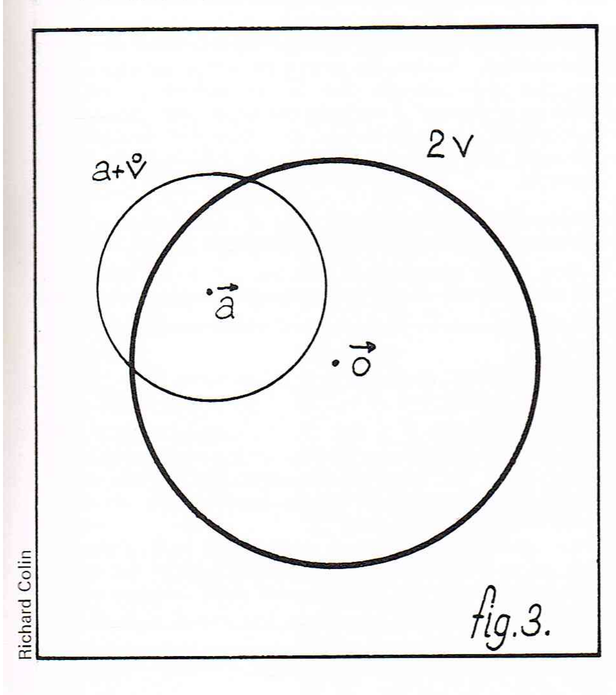
Espaces vectoriels topologiques de dimension finie

Dans le cas des espaces de dimension finie, tout devient notablement plus simple : il existe alors une seule topologie compatible avec la structure vectorielle et séparée (voir *Topologie*). Cette topologie sera dite « topologie canonique » ; elle se définit à l'aide d'une norme.

Le théorème suivant (dû à F. Riesz) établit une caractérisation des espaces de dimension finie (dans l'espace vectoriel de dimension finie $\mathbb{R}^n$, il y a identité entre les parties compactes et les parties fermées et bornées — ce qui n'est plus le cas en dimension infinie).

Théorème : *pour qu'un espace vectoriel topologique* E *soit de dimension finie, il faut et il suffit qu'il soit localement compact.*

Il convient de noter que ce résultat établit une équivalence entre une propriété de nature algébrique (dimension finie) et une propriété de nature topologique (la locale compacité). Donnons les grandes lignes de la démonstration (*fig.* 3) : soit V un voisinage compact de $\vec{0}$ dans E. 2 V possède, bien sûr, les mêmes propriétés. Soit $\vec{a}$ un élément quelconque de l'ensemble 2 V ; $\vec{a} + \mathring{V}$ est un *voisinage ouvert* de $\vec{a}$; en outre, lorsque $\vec{a}$ parcourt 2 V, les $\vec{a} + \mathring{V}$ constituent un recouvrement *ouvert* de 2 V :



$$2\,V \subset \bigsqcup_{\vec{a} \in 2V} (\vec{a} + \mathring{V})$$

D'après l'hypothèse de compacité, on peut en extraire un recouvrement fini (voir *Topologie*), c'est-à-dire qu'il existe des éléments $\vec{a}_1, \vec{a}_2, ..., \vec{a}_n$ tels que :

$$2\,V \subset \bigsqcup_{i=1}^{n} (\vec{a}_i + V_0).$$

Soit M le sous-espace vectoriel engendré par les a_i ; il est de dimension finie (au plus n) et M + V recouvre 2 V. La démonstration s'achève en montrant qu'en réalité M coïncide avec l'espace tout entier E qui, du même coup, se trouve être de dimension finie.

Ce théorème donne une idée des difficultés que l'on peut rencontrer en analyse fonctionnelle : les espaces de fonctions étant de dimension infinie, ils ne sont pas localement compacts. Aussi bien, il sera plus délicat d'obtenir des convergences de suites de fonctions (puisqu'il y a « moins » de compacts...). On déduit de ce théorème que tout sous-espace vectoriel de dimension finie est nécessairement *fermé* dans E.

Espaces vectoriels semi-normés

Considérons l'ensemble F^E des fonctions de E dans F lorsque F est un *espace vectoriel normé*. Nous avons défini sur cet espace la topologie de la convergence simple dont nous avons vu qu'elle n'était pas *normable*. Nous allons généraliser la notion d'espace normé en définissant celle d'*espace semi-normé*. Nous montrons que l'espace vectoriel topologique F^E est un espace vectoriel semi-normé.

Définition et propriétés élémentaires

Une semi-norme sur un espace vectoriel E est une fonction p : E → $\mathbb{R}^+$ qui possède les trois propriétés suivantes :

(5) $p(\vec{x}) \geqslant 0 ; p^r(\vec{0}) = 0 \qquad \forall \vec{x} \in E$

(6) $p(\lambda \cdot \vec{x}) = |\lambda|\, p(\vec{x}) \qquad \forall \vec{x} \in E, \lambda \in \mathbb{R}$

(7) $p(\vec{x} + \vec{y}) \leqslant p(\vec{x}) + p(\vec{y}) \qquad \forall \vec{x}, \vec{y} \in E.$

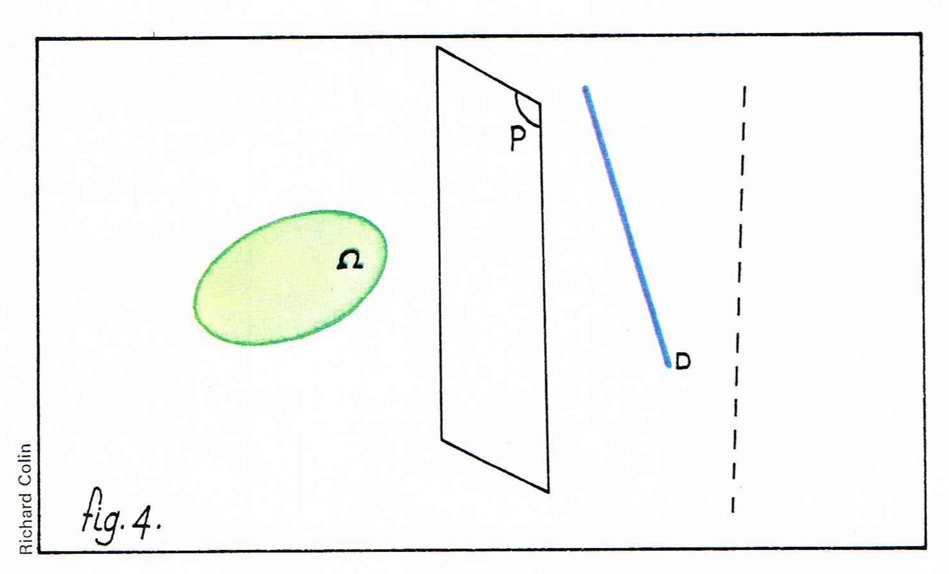
Seule la première de ces propriétés distingue une semi-norme d'une norme : en effet, il est possible d'avoir $p(\vec{x}) = 0$ avec $\vec{x} \neq \vec{0}$. E sera dit semi-normé s'il est muni d'une famille de semi-normes $(p_i)_{i \in I}$ (que l'on notera $\| \ \|_i$). Grâce à la donnée de cette famille, on peut munir E d'une topologie : on appellera *semi-boule ouverte* $B_i(\vec{a}, R)$ de centre $\vec{a}$ et de rayon R → 0 l'ensemble $B_i(\vec{a}, R) = \{\vec{x} \in E \mid \|\vec{x} - \vec{a}\|_i < R\}$.

Ainsi donc, une partie 0 de E sera dite *ouverte* si, pour tout point $\vec{a}$ de 0, il est possible de trouver une semi-boule de centre $\vec{a}$ entièrement contenue dans 0 ;

$$\forall a \in 0 \ \exists i \in I \quad \text{et} \quad \varepsilon_i > 0 \quad \text{tels que} \quad B_i(\vec{a}, \varepsilon_i) \subset 0.$$

On montrera sans peine que les axiomes des ouverts sont bien satisfaits. Il est nécessaire de supposer pour

◄ ***Figure 3 :** voir dans le texte la démonstration du théorème de Riesz de dimension finie.*



▲ *Figure 4 : forme géométrique (théorème de séparation) du théorème de Hahn-Banach.*

cela que la *famille* des $|p|_{i\ (i \in I)}$ est *filtrante*, c'est-à-dire telle que, pour toute partie $J \subset I$, il existe k tel que $p_k \geqslant p_j\ \forall j \in J$. Il est toujours possible de se ramener à ce cas-là en ajoutant aux (p_i) les sommes finies $\sum_{j \in J} p_j$ (voir *Topologie*).

Nous ne considérons que des espaces semi-normés dont la *topologie* est *séparée :* pour cela, il est nécessaire d'ajouter la condition supplémentaire :

$$\forall \vec{x} \neq \vec{0} \quad \exists i \in I \quad \|\vec{x}\|_i \neq 0.$$

Dans un espace semi-normé, la convergence s'exprime de la manière suivante ; nous dirons que $\vec{x}_n \to \vec{x}$ si

$$\forall i \in I, \forall \varepsilon > 0 \quad \exists N\ (i, \varepsilon)$$

$$\text{tel que } n > N\ (i, \varepsilon) \quad \|\vec{x}_n - \vec{x}\|_i \leqslant \varepsilon$$

D'autre part, la continuité d'une application linéaire d'un espace normé dans un autre s'exprime sous la condition : $\|f\ (\vec{x})\|_F \leqslant k\ \|\vec{x}\|_E$ où k est une constante positive. Lorsque l'espace de départ E est semi-normé, cette condition se généralise de la manière suivante : pour que f soit continue, il faut et il suffit qu'il existe *une* semi-norme $\|\ \ \|_i$ et une constante k telles que :

$$\|f\ (\vec{x})\|_F \leqslant k\ \|\vec{x}\|_i.$$

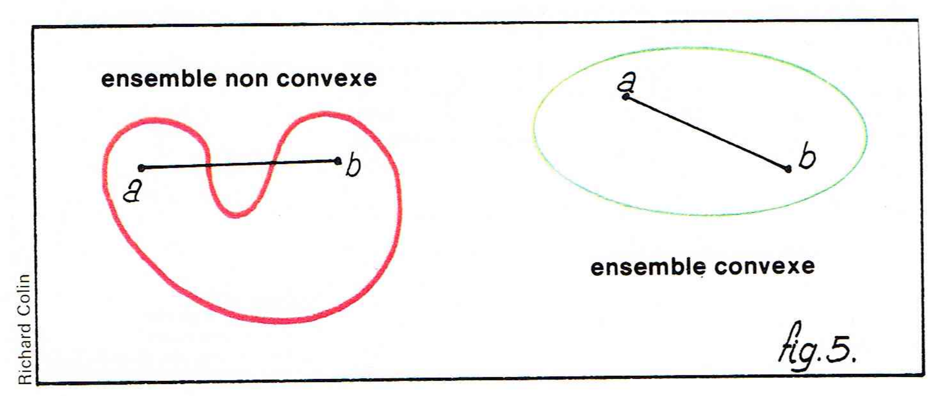
Enfin la topologie définie à l'aide d'une famille de semi-normes est compatible avec la structure vectorielle de l'espace : un *espace semi-normé est un espace vectoriel topologique*. D'autre part, on peut se restreindre aux seules semi-normes *continues* (en tant qu'applications de $E \to \mathbb{R}^+$) qui caractérisent la topologie.

Voyons maintenant comment ces notions permettent de décrire les topologies de certains espaces fonctionnels parmi les plus courants.

Exemples d'espaces vectoriels semi-normés

Considérons l'ensemble $\mathcal{C}\ (\mathbb{R}, \mathbb{R})$ des fonctions continues de $\mathbb{R}$ dans $\mathbb{R}$. Nous allons voir que cet espace fonctionnel peut être muni de deux topologies, chacune d'elles faisant de lui un espace semi-normé.

▼ *Figure 5 : ensemble convexe d'un espace vectoriel topologique.*



Convergence simple

On peut définir une famille de semi-normes de la manière suivante :

$$\|f\|_x = |f\ (x)| \quad x \in \mathbb{R}.$$

On vérifie que les $\|\ \ \|_x$ constituent une famille de semi-normes et qu'en outre la topologie qu'elle définit est séparée :

$$\|f\|_x = 0\ \forall x \Rightarrow |f\ (x)| = 0\ \forall x \Rightarrow f \equiv 0.$$

Dire enfin que la suite f_n de $\mathcal{C}\ (\mathbb{R}, \mathbb{R})$ converge vers f par la topologie des semi-normes, c'est dire que :

$$\|f_n - f\|_x \to 0\ \forall x \text{ ou encore que } f_n\ (x) \to f\ (x)$$

c'est exactement dire que f_n converge simplement vers f.

Pour que la famille de semi-normes soit filtrante, il est nécessaire de lui adjoindre d'autres semi-normes construites à l'aide des parties finies de $\mathbb{R}$: ainsi donc, soit $\mathcal{F}$ l'ensemble des parties finies de $\mathbb{R}$:

$$\mathcal{F} = \{A \subset \mathbb{R} \mid \text{Card } A < \infty\}$$

la famille de semi-normes qui définit la topologie de la convergence simple est donc la suivante :

$$\|f\|_A = \sup_{x \in A} |f\ (x)| \quad A \in \mathcal{F}.$$

Convergence compacte

Soit $\mathcal{K}$ l'ensemble des parties compactes de $\mathbb{R}$. Pour tout compact $K \in \mathcal{K}$, on définit :

$$\|f\|_K = \sup_{x \in K} |f\ (x)|.$$

Comme f est continue et K compact, cette expression a un sens. Il s'agit d'une semi-norme. Lorsque K parcourt $\mathcal{K}$, ces semi-normes définissent sur $\mathcal{C}\ (\mathbb{R}, \mathbb{R})$ une autre structure semi-normée.

Ces deux exemples montrent que, sur un même espace, il est possible de définir deux topologies compatibles avec la structure d'espace vectoriel. Comme $\mathcal{F} \subset \mathcal{K}$, la topologie de la convergence compacte $\mathcal{T}_C$ est plus fine que la topologie de la convergence simple $\mathcal{T}_S$. Cela signifie notamment que :

— tout ouvert A pour $\mathcal{T}_S$ est ouvert pour $\mathcal{T}_C$;

— une suite de fonctions f_n qui converge pour la topologie $\mathcal{T}_C$ converge également pour la topologie $\mathcal{T}_S$.

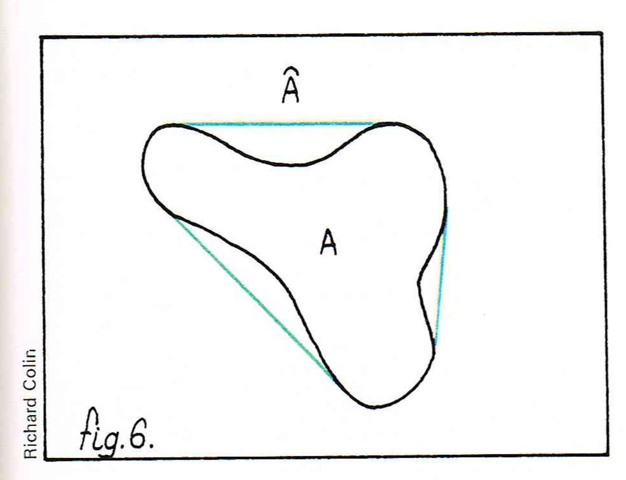
Espaces localement convexes - Théorème de Hahn-Banach

Nous introduisons ici la notion de convexité. Nous montrons en premier lieu qu'*un espace convexe est toujours semi-normable (et réciproquement)*. En second lieu, nous établirons le théorème de Hahn-Banach qui peut s'énoncer sous deux formes : une forme analytique (il s'agit d'un théorème de prolongement) et une forme géométrique (théorème de séparation). Sous sa forme géométrique la plus connue (*fig.* 4), il exprime l'idée intuitive selon laquelle, dans $\mathbb{R}^3$ par exemple, il est possible de « séparer » un *convexe* Ω et une droite D ne rencontrant pas Ω au moyen d'un plan P — la convexité est ainsi une notion essentielle au théorème de Hahn-Banach.

Ensembles convexes d'un espace vectoriel topologique

Soit C une partie ($\neq \varnothing$) d'un e. v. t. E ; on dira que C est convexe s'il vérifie la propriété suivante : quels que soient les points $\vec{a}_1, \vec{a}_2, ..., \vec{a}_m$ de C et les nombres $\lambda_1, ..., \lambda_m$ positifs et tels que $\sum_{i=1}^{m} \lambda_i = 1$, le point $\vec{a} = \sum_{i=1}^{m} \lambda_i \vec{a}_i$ appartient encore à C (*fig.* 5). On démontre qu'il suffit que cette propriété soit vérifiée pour deux points seulement. Dans $\mathbb{R}^n$, géométriquement, cela signifie que le segment joignant deux points quelconques de C est entièrement contenu dans C.

Considérons une partie A quelconque de E ; l'intersection de tous les convexes qui contiennent A est encore un convexe (*fig.* 6) ; c'est le plus petit convexe contenant A, et on l'appelle *enveloppe convexe de A (notée $\widehat{A}$)* ; on démontre que



◀ *Figure 6 : on appelle* **enveloppe convexe** *de A (notée $\hat{A}$) le plus petit convexe contenant A.*

$$\hat{A} = \left\{ \sum_{i=1}^{m} \lambda_i \vec{a_i} \mid \vec{a_i} \in A ; m \in \mathbb{N} \quad \lambda_i \geqslant 0 \quad \sum_{i=1}^{m} \lambda_i = 1 \right\}$$

Lorsque E est un espace vectoriel topologique, on peut établir quelques résultats, notamment : si C est un convexe, l'adhérence $\bar{C}$ et l'intérieur $\mathring{C}$ sont également convexes.

Espaces vectoriels topologiques localement convexes

Nous dirons que E est localement convexe s'il admet un système fondamental de voisinages *de l'origine* $\vec{O}$ *convexes*. Autrement dit, si tout voisinage V de $\vec{O}$ contient au moins un voisinage V′ de $\vec{O}$ *convexe*.

Énonçons le résultat : *pour qu'un espace vectoriel topologique soit localement convexe, il faut et il suffit qu'il soit semi-normable.*

Il est d'abord clair qu'un espace semi-normable est localement convexe : soit V un voisinage de $\vec{O}$, il contient au moins une demi-boule B_i (O, R), laquelle est, bien sûr, convexe. La réciproque est plus délicate à établir : elle consiste à construire une famille de semi-normes à partir d'un système fondamental de voisinages de $\vec{O}$ convexes et équilibrés (nous avons vu dans les *Espaces vectoriels topologiques* que l'on pouvait en trouver). Soit Ω un tel voisinage, on cherche alors une semi-norme p_Ω telle que Ω soit la semi-boule unité ouverte pour cette semi-norme

$$\Omega = \{\vec{x} \mid p_\Omega(\vec{x}) < 1\}$$

Si l'on parvient à définir une telle semi-norme, il est clair que l'on aura montré que la topologie de E est définie par la famille de semi-normes p_Ω où Ω parcourt l'ensemble des voisinages convexes, équilibrés et ouverts de $\vec{O}$.

A tout convexe C on peut associer une fonction j_C, appelée *jauge* du convexe C, définie par :

$$j_C(\vec{x}) = \inf\{k ; k \geqslant 0 \mid \vec{x} \in k\Omega\}$$

Cette définition signifie que $k \geqslant j_C(\vec{x}) \Leftrightarrow \frac{\vec{x}}{k} \in C$.

On démontre que, lorsque Ω est un ouvert convexe équilibré, la jauge associée est une *semi-norme continue* et qu'en outre $\Omega = \{x \mid j_\Omega(\vec{x}) < 1\}$. Autrement dit, la semi-norme cherchée n'est autre que la jauge de Ω (le fait que Ω soit équilibré est essentiel pour établir les axiomes de définition d'une semi-norme).

Forme analytique du théorème de Hahn-Banach

Sous sa forme analytique, le **théorème de Hahn-Banach** est un *théorème de prolongement*. Qu'est-ce à dire ? Soit $f : A \mapsto F$, on appellera prolongement de f à un ensemble $B \supset A$ une fonction $\tilde{f} : B \mapsto E$ qui coïncide avec f sur A. Autrement dit telle que $\tilde{f}(x) = f(x)\ \forall x \in A$. Dans tous les théorèmes de prolongement que l'on rencontre en analyse, on cherche un prolongement $\tilde{f}$ qui conserve certaines propriétés de f (par exemple continuité, linéarité, etc.) [*fig.* 7].

Le théorème de Hahn-Banach est un théorème de nature essentiellement vectorielle, et on n'y fait intervenir aucune hypothèse topologique. C'est précisément là que réside son intérêt puisqu'il peut être appliqué aux espaces de dimension infinie. Il n'en reste pas moins que l'on en déduira certaines propriétés topologiques dans le cas des espaces localement convexes.

La démonstration s'articule en deux étapes : dans un *lemme* préalable, on montre que l'on peut prolonger la fonction f à un espace possédant une dimension supplémentaire ; puis, grâce au fameux théorème de Zorn, on en déduit le prolongement à l'espace entier.

E est un espace vectoriel et p une *semi-norme*, F un sous-espace de E de *codimension* 1 (voir *Algèbre linéaire*), et f une forme linéaire définie sur F. ($f : F \to \mathbb{R}$) et telle que $f(\vec{x}) \leqslant p(\vec{x})\ \forall x \in F$; il existe alors au moins une forme linéaire $\tilde{f}$ définie sur E tout entier, prolongeant f et telle que $\forall x \in E\ \tilde{f}(\vec{x}) \leqslant p(\vec{x})$.

F est un hyperplan, autrement dit $E = F \oplus D$ où D est une droite (dimension 1).

Soit $\vec{x_1}$ un vecteur de D, donc $\notin$ F (*fig.* 8). Tout élément $\vec{x}$ de E s'écrit d'une manière unique sous la forme

$$\vec{x} = \lambda\vec{x}_1 + \vec{y}$$

où $\lambda \in \mathbb{R}$ et $\vec{y} \in F$. Une forme linéaire $\tilde{f}$ sur E prolongeant f s'exprime sous la forme générale suivante :

$$\tilde{f}(\vec{x}) = t\lambda + f(\vec{y}) \quad \text{où } \tilde{f}(\vec{x_1}) = t \in \mathbb{R}$$

peut être choisie arbitrairement. La démonstration s'achève en déterminant t afin que $\tilde{f}$ soit aussi majorée par p, c'est-à-dire telle que

$$f(\vec{y}) + \lambda t \leqslant p(\vec{y} + \lambda\vec{x_1}) \quad \forall\lambda \in \mathbb{R}, \vec{y} \in F$$

Donnons l'énoncé du **théorème de Hahn-Banach :**

Théorème : *soit E un espace vectoriel et p une semi-norme, F un sous-espace vectoriel quelconque de E, f une forme linéaire définie sur F et telle que*

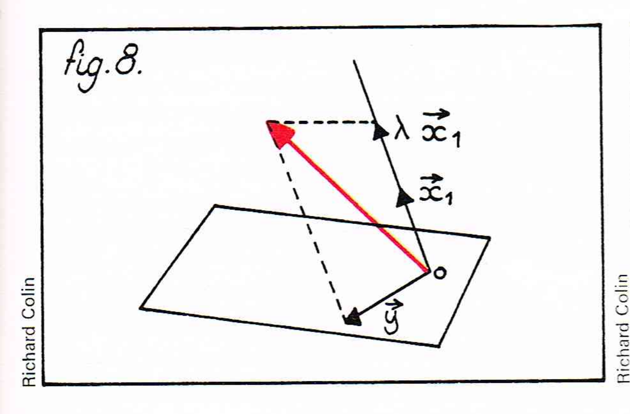
$$f(\vec{x}) \leqslant p(\vec{x}), \forall\vec{x} \in F.$$

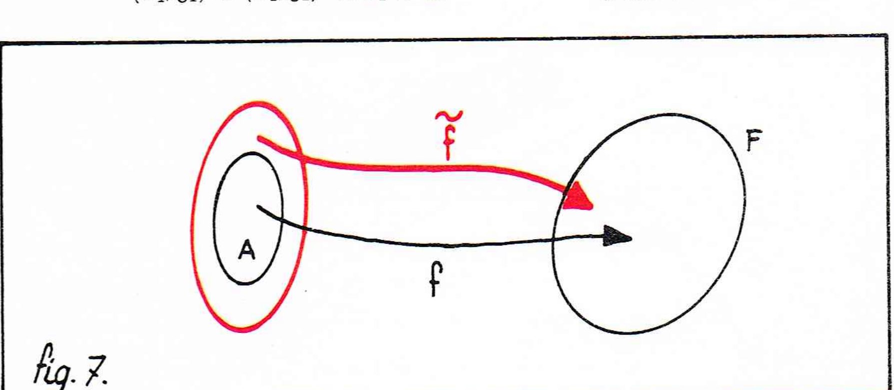
Il existe alors une forme linéaire (non nécessairement unique) définie sur E tout entier, prolongeant f *et telle que* $f(\vec{x}) \leqslant p(\vec{x})\quad \forall\vec{x} \in E$.

Utilisons donc le **théorème de Zorn :**

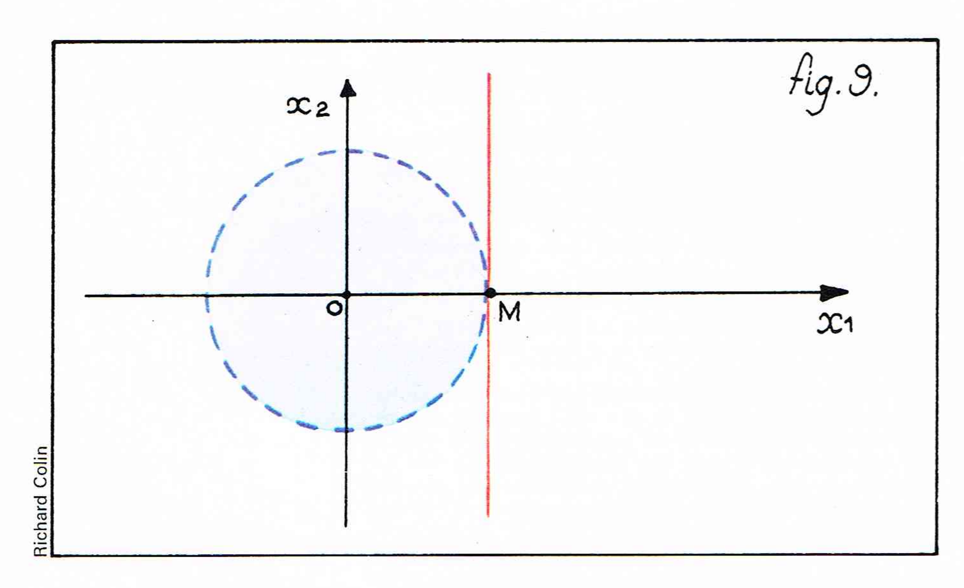
Théorème : *soit X l'ensemble des couples* (G, g) *où G est un sous-espace contenant F, et* g *un prolongement de* f *au sous-espace G et majoré par* p *sur G. On définit sur X une relation d'ordre* $\leqslant$ *de la manière suivante :*

$$(G_1, g_1) \leqslant (G_2, g_2) \quad \text{si } G_2 \supset G_1$$





▼ *A gauche, figure 8 : voir démonstration dans le texte.*
A droite, figure 7 : dans tous les théorèmes de prolongement que l'on rencontre en analyse, on cherche un prolongement $\tilde{f}$ qui conserve certaines propriétés de f.



▲ Figure 9 : exemple d'une application du théorème de Hahn-Banach sous sa forme géométrique.

et si g_2 est un prolongement de g_1. X est évidemment un ensemble ordonné et non vide car (F, f) $\in$ X.

On montre sans peine que X est un *ensemble inductif* (voir *Ensembles*).

Le théorème de Zorn conclut à l'existence d'un élément maximal $(E_1, \tilde{f})$. Utilisons le lemme précédent et montrons que $E_1 = E$: supposons que E_1 soit $\neq$ E, alors il existe $\vec{x}_1 \notin E_1$. Soit E_2 le sous-espace engendré par E_1 et $\vec{x}_1$. E_1 est un hyperplan de E_2 ; d'après le lemme, on pourrait prolonger $\tilde{f}$ à E_2, et, de ce fait, $(E_1, \tilde{f})$ ne serait plus maximal.

Remarque : dans l'énoncé du théorème, on peut remplacer l'inégalité $f(x) \leqslant p(x)$ par l'inégalité

$$|f(x)| \leqslant p(x).$$

Ces deux inégalités sont équivalentes : en effet,

$$f(-x) \leqslant p(-x) = p(x)$$

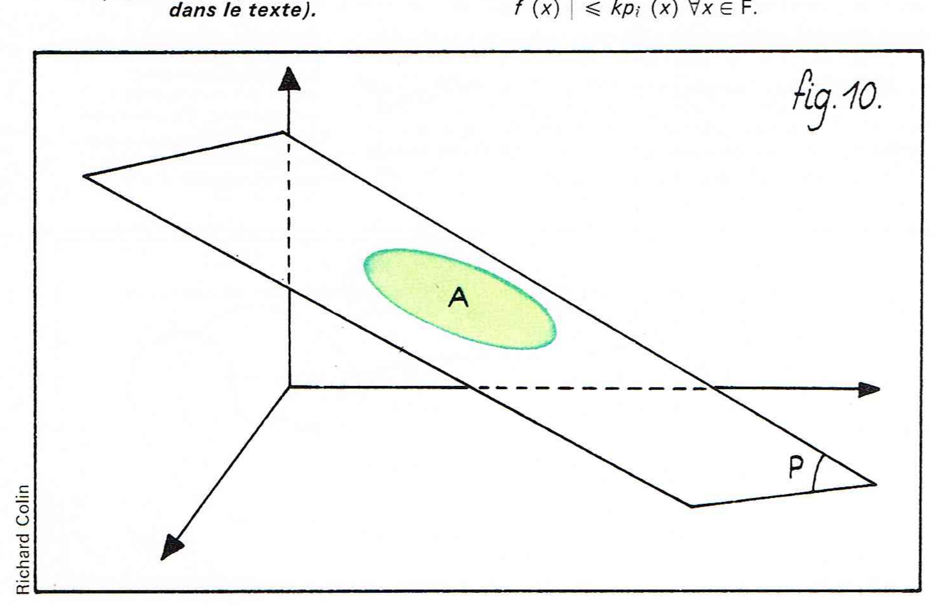
(p étant une semi-norme), soit $f(x) \geqslant -p(x)$, c'est-à-dire $|f(x)| \leqslant p(x)\ \forall x$.

Examinons ce que l'on peut déduire de ce théorème lorsque E est un espace vectoriel topologique : si l'application linéaire f et si la semi-norme p sont continues, alors le prolongement $\tilde{f}$ de f est lui aussi continu.

De cette première conséquence, on peut tirer le résultat suivant : *si E est localement convexe, F étant un sous-espace de E,* f *une forme linéaire continue sur F, alors il existe une forme linéaire continue $\tilde{f}$ sur E prolongeant* f : E est en effet alors semi-normable. Dire que f est continue sur F signifie qu'il existe une semi-norme (continue) p_i et une constante $k \geqslant 0$ telles que

$$|f(x)| \leqslant kp_i(x)\ \forall x \in F.$$

▼ Figure 10 : notion d'intérieur relatif (voir développement dans le texte).



Il existe donc un prolongement $\tilde{f}$ continu satisfaisant à la même inégalité. L'hypothèse de continuité est essentielle à la démonstration : elle ne suffit pas seulement à assurer la continuité de $\tilde{f}$, mais elle est tout à fait indispensable pour en établir l'existence même.

Remarque : le théorème de Hahn-Banach est également vrai lorsque p est seulement une *sous-norme*. Une sous-norme, comme une semi-norme, vérifie la semi-positivité, ainsi que l'inégalité de convexité (propriétés **(5)** et **(7)** de la définition du paragraphe : *Définitions et propriétés élémentaires*). Par contre, la propriété **(6)** n'est exigée que pour les homothéties de rapport > 0 :

$$p(\lambda\vec{x}) = \lambda p(\vec{x}) \quad \forall\lambda \geqslant 0.$$

On montre que la *jauge* d'un convexe ouvert Ω est une sous-norme ; elle est en plus une semi-norme si et seulement si Ω est *équilibré*. De cette expression plus générale du théorème de Hahn-Banach, l'on peut déduire sa forme géométrique.

Forme géométrique du théorème de Hahn-Banach

On montre en algèbre qu'un hyperplan H de E (sous-espace de codimension 1) peut toujours se définir comme le *noyau* d'une certaine forme linéaire f :

$$H = \{\vec{x} \in E \mid f(\vec{x}) = 0\}.$$

On démontre alors que H est fermé ou dense dans E suivant que f est continue ou non. De même, un *hyperplan affine* Ⓗ (c'est-à-dire le translaté d'un hyperplan H) se définit à l'aide d'une forme linéaire f et d'un nombre c de la manière suivante : Ⓗ $= \{\vec{x} \mid f(\vec{x}) = c\}$; Ⓗ sera fermé ou dense dans E suivant que f est continue ou non. En outre, on établit qu'il existe une seule forme linéaire f telle que Ⓗ $= \{\vec{x} \mid f(\vec{x}) = 1\}$. Ces quelques résultats liminaires vont nous permettre de démontrer la forme géométrique du **théorème de Hahn-Banach :**

Théorème : *soit E un espace vectoriel topologique, Ω un ouvert non vide, M un sous-espace affine non vide ne rencontrant pas Ω. Alors il existe au moins un hyperplan fermé affine* Ⓗ *contenant M et ne rencontrant pas non plus Ω.*

Nous illustrerons l'énoncé et la démonstration du théorème à l'aide de l'exemple géométrique suivant : prenons pour E le plan $\mathbb{R}^2$, pour Ω le *disque ouvert* de rayon 1 et pour M le point de coordonnées (1, 0). $\Omega = \{\vec{x} \mid \|\vec{x}\| < 1\}$, donc $\{M\} \cap \Omega = \varnothing$. M est un sous-espace affine (de dimension 0 !). Le théorème permet de conclure à l'existence d'une droite tangente à Ω au point M...

Démonstration du théorème (fig. 9) : Ω étant non vide, il est toujours possible de se ramener, par translation, au cas où $\vec{0} \in \Omega$; soit p la jauge de Ω (laquelle est une *sous-norme*) ; elle est de plus continue (dans l'exemple donné, p est la norme euclidienne). Soit F le sous-espace vectoriel engendré par M. Comme $\vec{0} \notin M$, M est un sous-espace affine et non vectoriel. M est donc strictement inclus dans F (dans l'exemple, F est l'axe des x_1), et de plus M est un *hyperplan affine de* F. Il existe donc une forme linéaire unique sur F telle que :

$$M = \{\vec{x} \mid f(\vec{x}) = 1\}$$

(dans l'exemple, $f : (0, x_1) \mapsto x_1$). Démontrons que $\vec{x} \in F \Rightarrow f(\vec{x}) \leqslant p(\vec{x})$. Soit $\vec{x} \in F$; deux cas peuvent se présenter : ou bien $f(\vec{x})$ est < 0, et alors $f(\vec{x}) \leqslant p(\vec{x})$ puisque p est $\geqslant 0$, ou bien $f(x)$ est > 0 ; il existe alors un nombre $\lambda = f(\vec{x})$ tel que $f(\lambda\vec{x}) = 1$, soit $\lambda\vec{x} \in M$, d'où $\lambda\vec{x} \notin \Omega$. C'est-à-dire $p(\lambda\vec{x}) \geqslant 1 = f(\lambda\vec{x})$, d'où l'on conclut $p(\vec{x}) \geqslant f(\vec{x})$, dans ce cas aussi. Finalement $f(\vec{x}) \leqslant p(\vec{x}) \quad \forall\vec{x} \in F$.

D'après le théorème de Hahn-Banach sous sa forme analytique, il existe une forme linéaire $\tilde{f}$ sur E, prolongeant f et telle que : $\tilde{f}(\vec{x}) \leqslant p(\vec{x})\ \forall\vec{x} \in E$ [dans l'exemple, $\tilde{f}$ est l'application $(x_1, x_2) \mapsto x_1$]. L'équation $f(\vec{x}) = 1$ définit un hyperplan affine fermé H ($\tilde{f}$ est continue) qui contient M.

Si $\vec{x} \in \Omega$, $\tilde{f}(\vec{x}) \leqslant p(\vec{x}) < 1$, alors $\vec{x} \notin H$; donc Ω ne

rencontre pas non plus H (dans l'exemple, H est la droite d'équation $x_1 = 1$, c'est-à-dire la tangente en M au cercle de centre O et de rayon 1).

Ce théorème permet d'obtenir de nombreux résultats en dimension finie, très utiles en optimisation linéaire et convexe. Donnons ici deux résultats :

— *Soit C_1 et C_2 deux convexes non vides de $\mathbb{R}^n$. Pour qu'il existe un hyperplan séparant C_1 et C_2, il faut et il suffit que les intérieurs relatifs de C_1 et de C_2 soient disjoints.*

Précisons la notion d'intérieur relatif (*fig.* 10) : on appelle intérieur relatif d'une partie A de $\mathbb{R}^n$ l'ensemble Ri (A), intérieur de A par rapport au sous-espace affine engendré par A. Exemple : A est un disque dans $\mathbb{R}^3$; il est clair que $\mathring{A}$ est $\varnothing$. Par contre, l'intérieur relatif de A est l'intérieur de A par rapport au plan p : c'est donc l'ensemble des points de A non situés sur le cercle.

Ce théorème prouve donc l'existence d'un hyperplan H d'équation $f(x) = b$ tel que :

$f(\vec{x_1}) \leqslant b \leqslant f(\vec{x_2}) \quad \forall \vec{x_1} \in C_1 \quad \forall \vec{x_2} \in C_2$ (noter les inégalités au sens large).

— *Soit C un convexe non vide fermé, et K un convexe compact non vide de $\mathbb{R}^n$ disjoints, alors il existe un hyperplan H d'équation* $f(\vec{x}) = b$ *qui les sépare strictement*

$$f(\vec{x}) < b < f(\vec{k}) \quad \forall \vec{x} \in C, \quad \forall \vec{k} \in K.$$

La dualité dans les espaces normés

L'idée de *dualité* est présente dans un grand nombre de résultats aussi bien d'algèbre que d'analyse, même si elle n'est pas toujours explicitée. Depuis longtemps, les mathématiciens avaient conscience que certains problèmes pouvaient susciter deux approches étroitement liées l'une à l'autre, deux approches en *« dualité »* dans un sens très intuitif qu'ils ont cherché à formaliser : il faudra attendre le XXe siècle et les travaux de Banach pour qu'elles reçoivent une formalisation rigoureuse.

Ainsi, une courbe plane peut être considérée sous deux aspects différents :

(a) soit comme l'ensemble de ses points liés par une équation $y = f(x)$;

(b) soit comme l'enveloppe de ses tangentes.

Deux points de vue différents, certes, mais que l'on sent doués d'une évidente complémentarité : suivant la nature du problème que l'on souhaite résoudre, on privilégiera un point de vue particulier en espérant que la résolution en sera plus aisée. Soit à résoudre, par exemple, le problème suivant (ô combien classique !) : « *trouver les extrema de la fonction $y = f(x)$* ». L'approche **(a)** se révèle n'être d'aucun intérêt puisqu'elle obligerait à calculer toutes les valeurs prises par la fonction afin de trouver celles qui fournissent les extrema (!) ; l'approche **(b)** est, à l'évidence, la plus féconde puisqu'elle stipule que les extrema correspondent aux points où la tangente est horizontale.

Un tel exemple montre l'intérêt qui peut exister à développer la notion de dualité (déjà évoquée en *Algèbre linéaire*) dans les espaces de fonctions. Appliquée aux espaces vectoriels topologiques, elle fournit un cadre propice à la résolution de certains problèmes d'analyse fonctionnelle. En outre, elle permet d'expliciter les fondements de la théorie de la mesure et de la théorie des distributions. Pour plus de simplicité, nous restreindrons notre étude aux *espaces normés*.

Dual topologique d'un espace normé E

On appelle *dual topologique de E* l'ensemble E′ de formes linéaires continues sur E. E′ est bien évidemment un espace vectoriel. Soit $\overleftarrow{e}'$ un élément de E′ ; on convient de noter $\langle \overleftarrow{e}', \vec{e} \rangle$ la valeur prise par la forme linéaire e' au point $\vec{e}$ [ou $\overleftarrow{e}'(\vec{e})$]. Cette notation met en évidence la symétrie qui existe entre l'espace E (le *primal*) et son *dual* E′. L'application de E × E′ ↦ $\mathbb{R}$ définie par

$$(\overleftarrow{e}' \vec{e}) \mapsto \langle \overleftarrow{e}', \vec{e} \rangle$$

est une application bilinéaire que l'on appelle également le *produit scalaire* de $\vec{e} \in$ E et de $\overleftarrow{e}' \in$ E par analogie avec l'exemple géométrique suivant :

Exemple : cas de la dimension finie $\mathbb{R}^n$

Le dual $\mathbb{R}^{n*}$ de $\mathbb{R}^n$ est l'ensemble des *vecteurs lignes* $\mathbb{R}^{n*} = \{\overleftarrow{a} = (a_1, ..., a_n) \mid a_i \in \mathbb{R}\}$. Soit $\vec{x}$ un vecteur de $\mathbb{R}^n$, $x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$. La valeur prise par la forme linéaire $\overleftarrow{a}$ au point $\vec{x}$ est donnée par :

$$\langle \overleftarrow{a}, \vec{x} \rangle = (a_1, ..., a_n) \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \sum_{i=1}^{n} a_i x_i.$$

Géométriquement, $\langle \overleftarrow{a}, \vec{x} \rangle$ est le produit scalaire des vecteurs $\vec{a}^t$ et $\vec{x}$.

En dimension finie, toutes les applications linéaires sont continues ; il y a donc coïncidence entre le *dual algébrique* (ensemble de toutes les formes linéaires, continues ou non) et le *dual topologique*. Cette coïncidence n'existe plus en dimension infinie. E′ étant un espace vectoriel, il est naturel de chercher à en faire un *espace vectoriel topologique* ; grâce à la relation de dualité qui le lie à E, il est possible de le munir de deux topologies différentes : la *topologie forte* et la *topologie faible*.

— *Topologie forte du dual :* on appelle topologie forte de E′ la topologie de la convergence uniforme sur les parties bornées ; cette topologie est définie à l'aide de la norme :

$$\| \overleftarrow{e}' \|_{E'} = \sup_{\|\vec{e}\|_E < 1} |\langle \overleftarrow{e}', \vec{e} \rangle|$$

Cette expression a un sens unique puisque $\overleftarrow{e}'$ est une forme linéaire continue (c'est-à-dire qu'il existe A > 0 tel que $|\langle \overleftarrow{e}', \vec{e} \rangle| \leqslant A \, \| \vec{e} \|$). L'espace E′ est un espace *normé complet* (puisque $\mathbb{R}$ l'est) ; c'est donc un *espace de Banach*.

— *Topologie faible du dual :* on appelle topologie faible de E′ la topologie de la convergence simple. Il s'agit d'une topologie semi-normable qui, ainsi que nous l'avons déjà vu, est définie par la famille de semi-normes $p_A(\overleftarrow{e}') = \sup_{\vec{e} \in A} |\langle \overleftarrow{e}', \vec{e} \rangle|$ où A est une partie finie de E.

C'est une topologie séparée ; en outre, c'est la moins finie des topologies compatibles avec la structure vectorielle et qui rendent continues les formes linéaires sur E′ $\overleftarrow{e}' \mapsto \langle \overleftarrow{e}', \vec{e} \rangle$ pour tout $\vec{e} \in$ E. Il est clair que la topologie forte est plus fine que la topologie faible.

Les espaces E et E′ sont doués d'une certaine symétrie au niveau de leurs structures vectorielles. Nous allons donc compléter cette symétrie en définissant une topologie affaiblie sur l'espace E qui sera l'analogue de la topologie faible de E′. L'espace E aura donc lui aussi deux topologies : la *topologie initiale* (ou topologie forte du primal) et cette *topologie affaiblie*.

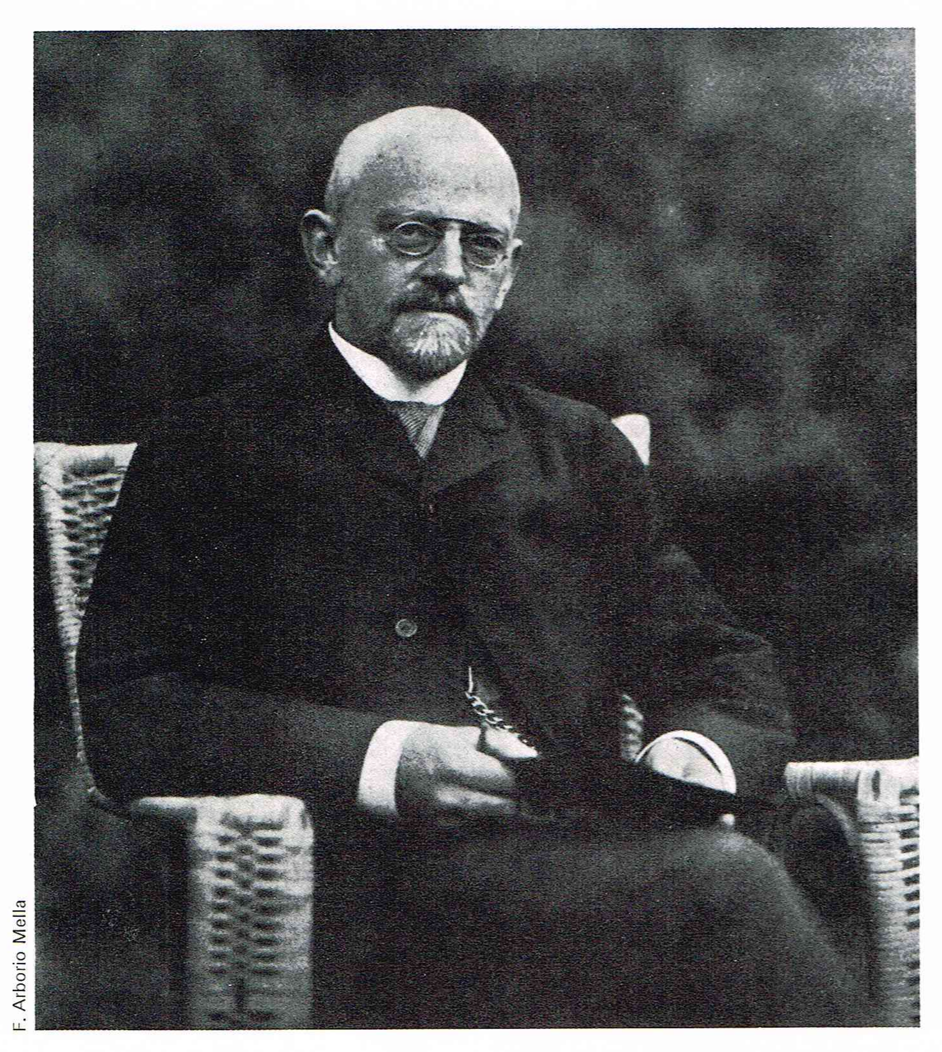
Topologie affaiblie du primal

On appelle *topologie affaiblie* de E la topologie définie par la famille de semi-normes $p_A(\vec{e}) = \sup_{\overleftarrow{e}' \in A} |\langle \overleftarrow{e}', \vec{e} \rangle|$ où A′ est une partie finie de E′. On note cette topologie σ (E, E′) : cette notation signifie qu'il s'agit d'une topologie *sur E* définie par des semi-normes construites à l'aide d'éléments *de E′* [on vérifiera ainsi que la topologie faible du dual E′ est la topologie σ (E′, E)]. On notera E_σ l'espace E muni de la topologie σ (E, E′). Il est possible alors de considérer le dual topologique de E_σ, c'est-à-dire l'ensemble des formes linéaires sur E, *continues pour la topologie affaiblie de* σ (E, E′).

La topologie forte de E (topologie initiale) étant plus fine que la topologie σ (E, E′), il est clair que toute forme linéaire *faiblement continue* (c'est-à-dire continue lorsque E est muni de la topologie affaiblie) sera *fortement continue*, autrement dit $(E_\sigma)' \subset$ E′. Inversement, soit e' une forme linéaire fortement continue, élément de E′ ; elle est, à l'évidence, faiblement continue puisqu'on a l'inégalité :

$$|\langle \overleftarrow{e}', \vec{e} \rangle| \leqslant p_{\overleftarrow{e}'}(\vec{e}) \quad \forall \vec{e} \in E \quad \text{où } p_{\overleftarrow{e}'}(\vec{e}) = |\langle \overleftarrow{e}', \vec{e} \rangle| \; (!)$$

Ainsi donc, $(E_\sigma)' =$ E′. La topologie σ (E, E′) est ainsi la topologie la moins fine compatible avec la structure vectorielle et pour laquelle E′ est encore le dual topologique de E.



F. Arborio Mella

▲ *Le mathématicien allemand* David Hilbert *(1862-1943).*

Dualité faible entre E et E′ : soit $\vec{e}$ un élément de E; l'application $\tilde{e} : \overleftarrow{e}' \in E' \mapsto \langle \overleftarrow{e}', \vec{e} \rangle$ est une forme linéaire *sur E′*, continue pour la topologie faible σ (E′, E). C'est donc un élément du dual topologique de l'espace E_σ. En identifiant la forme linéaire $\tilde{e}$ et l'élément $\vec{e}$ (voir *Algèbre linéaire*), on peut écrire $(E_\sigma) \supset E$. Grâce à la particularité de la topologie faible du dual, on montre que cette inclusion est en réalité une *égalité*. Ce résultat illustre bien la symétrie parfaite qui existe entre E_σ et E'_σ pour les topologies σ (E, E′) et σ (E′, E) : *chacun est le dual de l'autre, et la topologie affaiblie de l'un est la topologie faible du dual de l'autre.* Cette belle symétrie n'existe plus lorsque l'on considère les *topologies fortes* de E et de E′. Ainsi donc, dans le cas général, *E est inclus seulement dans le dual fort de E′*. C'est cette question que nous abordons maintenant.

Bidual : espaces réflexifs

Munissons donc E′ de la topologie forte; son dual topologique se note E″ et s'appelle le *bidual de E* et est lui-même muni de sa topologie forte (que l'on définit en utilisant la dualité entre E′ et E″. Nous identifierons E à un sous-espace de E″ par le procédé évoqué ci-dessus (au surplus E, étant normé, $\| \vec{e} \|_E = \| \tilde{e} \|_{E''}$

$$E \to E' \to E'' \ldots$$

On dira que E *est réflexif s'il est identique à son bidual : $E \equiv E''$.*

D'ores et déjà, nous pouvons remarquer une condition nécessaire de réflexivité : l'espace E doit être un espace de Banach; mais cette condition n'est pas suffisante. En réalité, une condition nécessaire et suffisante est donnée par le théorème de Banach.

Théorème : *pour qu'un espace normé E soit réflexif, il faut et il suffit que sa boule unité soit faiblement compacte.*

Ce résultat est à rapprocher du théorème de F. Riesz qui fournit une caractérisation des espaces de dimension finie par la compacité (forte) de la boule unité. En dimension finie, la topologie forte et la topologie faible coïncident, et la réflexivité est toujours réalisée. Le théorème de Banach est ainsi une généralisation du théorème de Riesz. Il illustre l'intérêt des espaces réflexifs : que la boule unité soit faiblement compacte (ce qui entraîne d'ailleurs que l'espace est complet) signifie que l'on pourra établir certaines convergences (faibles).

Nous démontrerons seulement un résultat liminaire au théorème de Banach qui situe bien la nature du problème :

Proposition : *soit E un espace normé, la boule unité de son dual E′ est faiblement compacte.*

La topologie faible sur E′ est celle de la convergence simple. Autrement dit, il s'agit de la topologie induite par l'espace $\mathbb{R}^E$ de toutes les fonctions sur E à valeurs réelles. Pour montrer que B′ est faiblement compacte, on démontrera que B′ est inclus dans une partie compacte de $\mathbb{R}^E$, et qu'en outre B′ est fermé dans $\mathbb{R}^E$. Il est clair que

$B' \subset \prod_{\vec{e} \in A} \overline{A(\vec{e})}$ où $A(\vec{e}) = \{\langle \overleftarrow{e}', \vec{e} \rangle \mid \overleftarrow{e}' \in B'\}$. $A(\vec{e})$ est un intervalle borné de $\mathbb{R}$: en effet,

$$\overleftarrow{e}' \in B' \Rightarrow |\langle \overleftarrow{e}', \vec{e} \rangle| \leqslant \|\vec{e}\|.$$

$A(\vec{e})$ est donc un compact de $\mathbb{R}$ et, d'après le théorème de Tychonoff (voir *Topologie*), l'ensemble $\prod \overline{A(e)}$, produit de compacts, est compact.

Reste à montrer que B′ est fermé dans $\mathbb{R}^E$: soit $\vec{x}$ et $\vec{y}$ des éléments de E, et λ et μ deux nombres réels. Pour $\vec{x}, \vec{y}, \lambda, \mu$ fixés, la fonction $\Phi_{\vec{x}, \vec{y}, \vec{\lambda}, \vec{\mu}} : \mathbb{R}^E \to \mathbb{R}^E$ définie par $f \mapsto f[\lambda \vec{x} + \mu \vec{y}] - \lambda f(\vec{x}) - \mu f(\vec{y})$ est évidemment continue (pour tout $z \in E$, l'application $f \mapsto f(z)$ est une projection de $\mathbb{R}^E$ sur un de ses facteurs). Donc l'ensemble des fonctions f vérifiant $\Phi_{\vec{x}, \vec{y}, \vec{\lambda}, \vec{\mu}}(f) = 0$, image réciproque du fermé $\{0\}$, est fermé dans $\mathbb{R}^E$. Soit E^* le *dual* algébrique de E. Il est clair que :

$$E^* = \bigcap_{\substack{\vec{x}, \vec{y} \in E, \\ \lambda, \mu \in \mathbb{R}}} \{f / \Phi_{\vec{x}, \vec{y}, \lambda, \mu}(f) = 0\}.$$

Autrement dit, E^*, étant une intersection de fermés, est fermé dans $\mathbb{R}^E$. Soit $\vec{x}$ enfin un élément de E. L'application Ψ_x de $\mathbb{R}^E$ dans $\mathbb{R}$ définie par $\Psi_x(f) = f(x) - \|x\|$ est non moins évidemment continue. Par conséquent, l'ensemble $A_x = \{f \mid \Psi_x(f) \leqslant 0\}$ est une partie fermée de $\mathbb{R}$ (comme image réciproque du fermé $\mathbb{R}^-$), et l'ensemble $A = \bigcap_{\vec{x} \in E} A_x$ l'est également. Il reste à remarquer que $B' = E^* \cap A$, et donc que B′ est fermé dans $\mathbb{R}^E$, c'est-à-dire faiblement fermé dans E′. Ainsi donc, *B′ est faiblement compact dans E′*. Ce résultat liminaire intervient dans la démonstration du théorème de Banach en remarquant que, si E est réflexif, alors il est le dual topologique de E′, et donc, en tant que dual, sa boule unité doit être relativement compacte.

Exemples d'espaces de Banach réflexifs : les espaces $L^p(\mathbb{R}^n; \mathbb{R})$.

Soit $L^p(\mathbb{R}^n; \mathbb{R})$ l'ensemble de fonction f de $\mathbb{R}^n \to \mathbb{R}$ p-intégrables (voir *Analyse*) $p \in \{1, 2, \ldots, N, \ldots\}$. L^p est un espace vectoriel normé pour la norme

$$\|f\|_p = \left(\int_{\mathbb{R}} |f(x)|^p\right)^{\frac{1}{p}}$$

On montre que, pour $p \geqslant 2$, le dual de L^p est l'espace L^q où p et q sont liés par l'égalité $\frac{1}{p} + \frac{1}{q} = 1$. Les espaces L^q sont donc tous réflexifs pour $p \geqslant e$. Par contre, pour $p = 1$, on montre que son dual est L^∞ (ensemble des fonctions presque partout bornées sur $\mathbb{R}^n$), mais le dual de L^∞ est strictement plus grand que L^1.

Espaces de Hilbert

Les espaces de Hilbert constituent un cas particulier des espaces de Banach réflexifs. Ils sont les espaces vectoriels topologiques les plus « proches » des espaces de dimension finie ($\mathbb{R}^n$ est d'ailleurs un espace de Hilbert), car ils présentent un certain nombre de qualités géométriques : il existe un isomorphisme canonique entre

l'espace et son dual topologique ; en outre, il est possible de « projeter » sur tout sous-espace fermé. Enfin, on généralise la notion de base en définissant une base hilbertienne, composée d'une infinité dénombrable d'éléments.

Forme bilinéaire

Soit un espace vectoriel, on appelle forme bilinéaire B une application de $E \times E \to \mathbb{R}$ *linéaire par rapport à chacune des variables :*

$$B(\lambda \vec{x}_1 + \mu \vec{x}_2, \vec{y}) = \lambda B(\vec{x}_1, \vec{y}) + \mu B(\vec{x}_2, \vec{y}) \quad \forall \vec{x}_1, \vec{x}_2, \vec{y} \in E$$
$$B(\vec{x}, \lambda \vec{y}_1 + \mu \vec{y}_2) = \lambda B(\vec{x}, \vec{y}_1) + \mu B(\vec{x}, \vec{y}_2) \quad \forall \vec{y}_1, \vec{y}_2, \vec{x} \in E \quad \forall \lambda, \mu \in \mathbb{R}.$$

Une forme bilinéaire est dite *positive* si

$$B(\vec{x}, \vec{x}) \geqslant 0 \quad \forall \vec{x} \in E.$$

Elle sera dite *définie positive* si, en outre, $B(\vec{x}, \vec{x}) > 0$ pour $\vec{x} \neq 0$. Enfin elle sera dite *symétrique* si

$$B(\vec{x}, \vec{y}) = B(\vec{y}, \vec{x}).$$

Si B est une forme bilinéaire symétrique et positive, on a l'**inégalité de Schwarz** :

$$|B(\vec{x}, \vec{y})| \leqslant B(\vec{x}, \vec{x})^{\frac{1}{2}} \cdot B(\vec{y}, \vec{y})^{\frac{1}{2}}.$$

Rappelons-en la démonstration : on peut écrire

$$B(\vec{x} + \lambda \vec{y}, \vec{x} + \lambda \vec{y}) = B(\vec{x}, \vec{x}) + 2\lambda B(\vec{x}, \vec{y}) + \lambda^2 B(\vec{y}, \vec{y})$$

avec $\lambda \in \mathbb{R}$. Dire que B est positive signifie que :

$$B(\vec{x} + \lambda \vec{y}, \vec{x} + \lambda \vec{y}) \geqslant 0 \quad \forall \lambda \in \mathbb{R}$$

soit que l'équation en λ :

$$\lambda^2 B(\vec{y}, \vec{y}) + 2\lambda B(\vec{x}, \vec{y}) + B(\vec{x}, \vec{x})$$

n'a pas de racine réelle, c'est-à-dire

$$\Delta = B^2(\vec{x}, \vec{y}) - B(\vec{x}, \vec{x}) \cdot B(\vec{y}, \vec{y}) \leqslant 0.$$

Espace de Hilbert

Soit H un espace vectoriel sur lequel on a pu définir une forme bilinéaire symétrique, définie positive (ou produit scalaire que l'on va noter $(\vec{x} \mid \vec{y})$). En utilisant l'inégalité de Schwarz, on montre que l'application $\vec{x} \to (\vec{x} \mid \vec{x})^{\frac{1}{2}}$ est une norme. Si H est *complet* pour la dite norme, on dira que H est un *espace de Hilbert.*

Exemples :

$\mathbb{R}^n$ est un espace de Hilbert pour le produit scalaire $(\vec{x} \mid \vec{y}) = \sum_{i=1}^{n} x_i y_i.$

L'espace $L^2(\mathbb{R}^n, \mathbb{R})$ des fonctions de $\mathbb{R}^n$ dans $\mathbb{R}$ de carré intégrable (telles que $\int_{\mathbb{R}^n} (f(x))^2 \, dx < +\infty$) est également un espace de Hilbert pour le produit scalaire : $(f \mid g) = \int_{\mathbb{R}^n} f(x) \cdot g(x) \, dx.$

Le théorème de projection

Dans un espace de Hilbert, on définit la notion de projection de manière analogue à ce que l'on réalise géométriquement dans $\mathbb{R}^n$. De nombreux problèmes d'analyse numérique (problèmes d'optimisation, équations aux dérivées partielles, etc.) se résolvent en caractérisant la solution comme la projection d'un certain élément d'un espace de Hilbert (espace L^2 ou *espace* dit *de Sobolev*) sur un sous-ensemble dûment choisi de l'espace. Le théorème s'énonce comme suit :

Théorème : *soit H un espace de Hilbert et F une partie fermée, convexe et non vide de H. Alors chaque point $\vec{a}$ de H a une projection et une seule sur F.*

La démonstration du théorème (*fig.* 11) s'appuie sur un résultat bien connu des géomètres sous le nom de « lemme de la médiane » : étant donné trois points $\vec{a}, \vec{b}, \vec{c}$ d'un espace de Hilbert, si m est le milieu de $(\vec{b}, \vec{c})$, on a l'égalité

$$\| \vec{a} - \vec{b} \|^2 + \| \vec{a} - \vec{c} \|^2 = 2 \| \vec{a} - \vec{m} \|^2 + \frac{1}{2} \| \vec{b} - \vec{c} \|^2.$$

Soit donc $\vec{a}$ un point de H. On appelle distance de a à F le nombre $d = \min_{\vec{x} \in F} \| \vec{x} - \vec{a} \|$. On peut trouver une suite $\vec{x}_1, \ldots, \vec{x}_n$ de points F telle que $d = \lim_{n \to +\infty} \| \vec{x}_n - \vec{a} \|$.

Il nous faut montrer que la suite des $\vec{x}_n$ converge vers un point $\vec{\alpha}$ de F ; la norme $\| \quad \|$ étant une application continue, $\vec{\alpha}$ sera la projection de $\vec{a}$ sur F.

Appliquons le lemme de la médiane aux trois points $\vec{a}$, $\vec{x}_m$ et $\vec{x}_n$ de F :

$$\frac{1}{2} \| \vec{x}_m - \vec{x}_n \|^2 = \| \vec{a} - \vec{x}_n \|^2 + \| \vec{a} - \vec{x}_m \|^2 - 2 \left\| \vec{a} - \frac{\vec{x}_m + \vec{x}_n}{2} \right\|^2.$$

Lorsque m et $n \to \infty$, $\| \vec{a} - \vec{x}_n \|^2$ et $\vec{a} - \vec{x}_m \|^2$ tendent vers d^2. En outre, F étant convexe $\frac{\vec{x}_m + \vec{x}_n}{2}$ appartient à F, et donc $\left\| \vec{a} - \frac{x_m + x_n}{2} \right\| \geqslant d^2$. Ainsi donc, le deuxième membre de l'égalité a une limite négative. Or cette limite ne pouvait être que $\geqslant 0$; il en résulte que

$$\lim_{n, m \to \infty} \| \vec{x}_n - \vec{x}_m \| = 0.$$

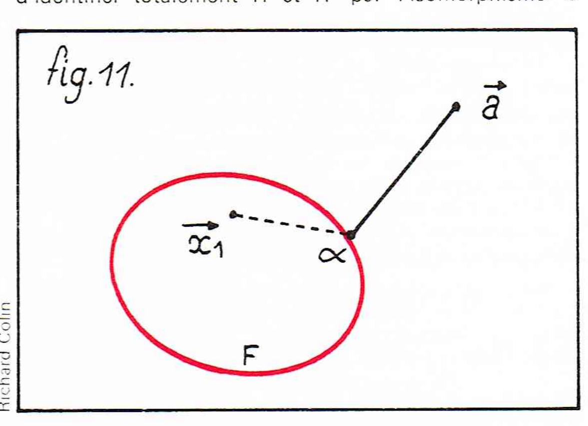
Autrement dit, la suite des $\vec{x}_n$ est une suite de Cauchy. Comme E est *complet*, cette suite converge vers un élément $\vec{\alpha}$ qui appartient encore à F puisque F est fermé dans H. Il est clair que $\| \vec{a} - \vec{\alpha} \|^2 = d^2$ et que cette limite est unique (l'unicité est due à la convexité de F).

Ce résultat, géométriquement évident, acquiert tout son intérêt lorsqu'on l'applique à des espaces de Hilbert de dimension infinie : en théorie des probabilités par exemple, une variable aléatoire X réelle se définit comme une application particulière (mesurable) d'un ensemble Ω, espace des « événements » dans $\mathbb{R}$; X est alors considéré comme un élément d'un espace fonctionnel $L^2(\Omega, \mathcal{B}, \mathbb{R})$ ($\mathcal{B}$ étant une *σ-algèbre* définie sur Ω). Soit H_X le sous-espace de $L^2(\Omega, \mathcal{B}, \mathbb{R})$ engendré par la variable X. On démontre :
$H_X = \{f \circ X \mid$ où f appartient à une certaine classe $\mathcal{L}$ de fonctions de $\mathbb{R} \to \mathbb{R}\}$. H_X est un sous-espace (donc convexe) et fermé de $L^2(\Omega, \mathcal{B}, \mathbb{R})$. Soit une autre variable aléatoire Y. On définit l'espérance conditionnelle de Y par X $\left(\text{notée } E\left(\frac{Y}{X}\right)\right)$ comme la *variable aléatoire, projection de Y sur H_X*.

Dualité dans les espaces de Hilbert

La particularité essentielle des espaces de Hilbert est qu'il existe un isomorphisme linéaire continu entre H et son dual topologique H' : soit $\vec{y} \in H$, on lui fait correspondre, canoniquement, la *forme linéaire* (continue) $U(\vec{y}) : \vec{x} \to (\vec{y} \mid \vec{x})$; $U(\vec{y})$ est un élément du dual H' pour tout $\vec{y}$. En reprenant la notation introduite au chapitre précédent, on écrit : $\langle U(\vec{y}), \vec{x} \rangle = (\vec{y} \mid \vec{x})$.

U est donc une application de H dans H' dont on montre qu'elle est inversible, linéaire, continue et qu'en outre, il s'agit d'une isométrie : c'est-à-dire qu'elle conserve les distances $\| U(\vec{y}) \|_{H'} = \| \vec{y} \|_H$. On conviendra donc d'identifier totalement H et H' par l'isomorphisme U



◀ ***Figure 11 :*** ***le théorème de la projection généralisé aux espaces de Hilbert, une propriété géométrique de l'espace euclidien $\mathbb{R}^3$.***

(ce qui, du même coup, revient à identifier le produit scalaire $(\cdot \mid \cdot)$ et l'opération de dualité $\langle \cdot, \cdot \rangle$. Il ressort de l'existence d'un tel isomorphisme qu'un espace de Hilbert est un espace réflexif. L'on peut donc en déduire que la boule unité de H est *faiblement compacte* (théorème de Banach).

Base hilbertienne d'un espace de Hilbert

Soit $(\vec{e_i})_{i \in I}$ une famille de vecteurs d'un espace de Hilbert H. On dira qu'elle constitue un système orthonormé si

$$\|\vec{e_i}\| = 1 \quad \forall i \in J$$
$$(\vec{e_i} \mid \vec{e_j}) = 0 \; i \neq j.$$

Les vecteurs $\vec{e_i}$ sont donc unitaires et deux à deux orthogonaux. Une telle famille sera dite *base hilbertienne* si, de surcroît, le sous-espace engendré par les $\vec{e_i}$ est *dense* dans H. Cela signifie que tout vecteur $\vec{x}$ de H va s'exprimer comme limite d'une série $\sum_{i \in I} x_i \vec{e_i}$ où une infinité au plus dénombrable des x_i est $\neq 0$.

En utilisant le théorème de Zorn, on démontre que *tout espace de Hilbert admet des bases hilbertiennes*, et, qui plus est, qu'elles sont toutes équipotentes : c'est-à-dire si $(\vec{e_i})_{i \in I}$ et $(\vec{f_j})_{j \in I}$ sont deux bases hilbertiennes, Card I = Card J.

Cette propriété montre combien la base hilbertienne est une généralisation de la base que l'on obtient en dimension finie; elle est d'un grand intérêt en analyse, ainsi qu'en témoignent les 2 exemples suivants.

Polynômes de Legendre

Soit $\mathcal{C}\,[-1, +1]$ l'ensemble des fonctions réelles continues sur $[-1, +1]$, et munissons-le du produit scalaire $(f|g) = \int_{-1}^{+1} f(x) \cdot g(x)\, dx$. L'espace $\mathcal{C}\,[-1, +1]$ muni d'un tel produit scalaire n'est pas un espace de Hilbert puisqu'il n'est pas complet (c'est un espace dit préhilbertien). Néanmoins, il est possible de le « plonger » dans un espace plus grand H, son « complété », dont on montre en intégration qu'il n'est autre que l'espace $L^2([-1, +1], dx)$.

Appelons polynôme de Legendre de degré $n \geqslant 0$ le polynôme :

$$X_n(x) = \frac{1}{2^n\, n!} \frac{d^n}{dx^n} [(x^2 - 1)^n].$$

On vérifie que la suite des polynômes $P_n = \dfrac{\sqrt{2n+1}}{\sqrt{2}} X_n$ est un système orthonormé de H, c'est-à-dire que

$$\int_{-1}^{+1} P_n(x) \cdot P_m(x)\, dx = \begin{cases} 0 & n \neq m \\ 1 & n = m. \end{cases}$$

Reste à montrer que c'est une base hilbertienne de H. D'abord les P_n engendrent tout l'espace des polynômes (ils sont indépendants puisque orthogonaux); ensuite l'ensemble des polynômes est dense dans $\mathcal{C}\,[-1, +1]$ pour la topologie de la convergence uniforme (ce résultat est connu sous le nom de **théorème de Stone-Weierstrass**), donc *a fortiori* pour la topologie préhilbertienne qui est moins fine. Enfin $\mathcal{C}\,[-1, +1]$ est dense dans H qui en est le complété.

Décomposition en série de Fourier

L'on considère ici des espaces de Hilbert sur le corps $\mathbb{C}$. La généralisation se fait sans peine, le produit scalaire $(\,|\,)$ étant assujetti à être une forme sesquilinéaire : elle doit vérifier : $B[\vec{x}, \lambda \vec{y}] = \bar{\lambda} B(\vec{x}, \vec{y})$ ou $\lambda \in \mathbb{C}$ et $\bar{\lambda}$ conjugué de λ...

Soit $\mathcal{T}$ l'ensemble des fonctions continues, périodiques sur $\mathbb{R}$ de période 1. On montre que la famille des fonctions $\{e^{2i\pi nx}\}_{n \in \mathbb{Z}}$ constitue une base hilbertienne de $\mathcal{T}$. Les « composantes » d'une fonction f sur cette base sont données par la formule bien connue :

$$C_n(f) = \int_0^1 f(x)\, e^{-2i\pi nx}\, dx \quad n \in \mathbb{Z}$$

et la série $\sum_{n \in \mathbb{Z}} C_n(f)\, e^{2i\pi nx}$ converge vers f dans H (H étant ici le complété de $\mathcal{T}$).

Les distributions

Historiquement, les distributions ont été introduites par les physiciens dans ce qu'ils appellent le calcul symbolique. Il s'agissait de concepts peu rigoureux mais qui possédaient l'avantage de simplifier notablement les calculs : ainsi, les physiciens étaient conduits à considérer la dérivée de fonctions discontinues auxquelles ils appliquaient les règles de calcul usuelles. Ce genre de pratiques était assurément fort suspect aux yeux des mathématiciens; encore fallait-il expliquer pourquoi elles réussissaient si bien aux physiciens. Il revient au mathématicien français Laurent Schwartz d'avoir élaboré une théorie mathématique nouvelle qui a permis de résoudre cette contradiction (1947) : en définissant une distribution comme un élément du dual d'un espace de fonctions particulier, il a donné la justification théorique du langage des physiciens. Nous présentons donc les distributions en montrant quels phénomènes physiques elles permettent de décrire.

Définition d'une distribution

Soit $\mathfrak{D}$ l'espace vectoriel des fonctions $\varphi : \mathbb{R}^n \to \mathbb{R}$ *indéfiniment dérivables et à support compact* (chaque fonction φ est nulle en dehors d'un compact K_φ que l'on appelle support de φ). Par exemple, la fonction $\theta : \mathbb{R} \to \mathbb{R}$ définie par :

$$\theta(x) = \begin{cases} 0 & \text{si } |x| > 1 \\ e^{\frac{-1}{1-x^2}} & \text{si } |x| < 1 \end{cases}$$

est un élément de $\mathfrak{D}$ (pour $n = 1$) : il faut vérifier que θ est dérivable partout (surtout en $+1$ et -1) et qu'en ces deux points *toutes* ses dérivées sont nulles. L'espace $\mathfrak{D}$ est un espace de fonctions que l'on peut munir d'une topologie adaptée à sa structure propre : nous définissons cette topologie en caractérisant la notion de convergence (cela nous suffira amplement ici; mais on montre que cette *topologie est semi-normable*) : on dira qu'une suite φ_j de $\mathfrak{D}$ converge vers une fonction φ de $\mathfrak{D}$ si :

— les supports des φ_j sont contenus dans un même ensemble borné, indépendant de j;

— les dérivées de tout ordre m des φ_j convergent uniformément vers les dérivées correspondantes de φ.

On appelle distribution T une forme linéaire continue sur l'espace vectoriel topologique $\mathfrak{D}$. Autrement dit, une distribution est un élément du dual topologique $\mathfrak{D}'$ de $\mathfrak{D}$.

Une distribution possède donc les propriétés suivantes :

$$\langle T, \varphi_1 + \varphi_2 \rangle = \langle T, \varphi_1 \rangle + \langle T, \varphi_2 \rangle$$
$$\langle T, \lambda \varphi \rangle \quad = \lambda \langle T, \varphi \rangle$$

si φ_j converge vers φ au sens de la topologie de $\mathfrak{D}$, les nombres $\langle T, \varphi_j \rangle$ convergent vers $\langle T, \varphi \rangle$.

Exemples

(a) Soit f une fonction de $\mathbb{R}^n \to \mathbb{R}$, intégrable pour la mesure de Lebesgue; elle définit une distribution T_f par :

$$\langle T_f, \varphi \rangle = \int_{\mathbb{R}^n} f(\vec{x})\, \varphi(\vec{x})\, dx_1, ..., dx_n.$$

Cette intégrale a bien un sens puisque la fonction φ est nulle en dehors d'un certain compact K_φ. Il s'agit bien d'une forme linéaire dont il nous reste à établir la continuité : soit φ_j une suite de fonctions convergentes vers φ dans $\mathfrak{D}$: soit K le compact contenant tous les supports des φ_j $|\langle T_f - \varphi_j \rangle - \langle T_f, \varphi \rangle| =$

$$\left| \int_K f(\vec{x}) \cdot \varphi_j(x) \cdot dx_1, ... dx_n - \int_K f(\vec{x}) \cdot \varphi(\vec{x})\, dx_1, ... dx_n \right|$$

$$\leqslant \left(\int_K |f(x)|\, dx_1, ... dx_n \right) . \operatorname*{Max}_{x \in K} \left| \varphi(\vec{x}) - \varphi_j(\vec{x}) \right|$$

Comme la convergence des φ_j est uniforme,

$$\operatorname*{Max}_{x \in K} |\varphi(\vec{x}) - \varphi_j(\vec{x})| \to 0$$

et ainsi $\langle T_f, \varphi_j \rangle \to \langle T_f, \varphi \rangle$. On remarque que, dans ce cas, T_f est la mesure de densité f. Nous conviendrons désormais d'identifier la distribution T_f et la fonction f : ainsi, nous dirons qu'une distribution est une fonction f si la valeur qu'elle prend pour toute fonction φ de $\mathfrak{D}$ est donnée par

$\langle T, \varphi\rangle \int_{\mathbb{R}^n} f(x)\, \varphi(x)\, dx_1, \ldots, dx_n.$

(b) Plus généralement, notons D un opérateur de dérivation partielle d'ordre quelconque en $x_1, \ldots, x_n$: la forme linéaire

$$\langle T, \varphi\rangle = \int_{\mathbb{R}^n} f(x)\, D\varphi(x)\, dx_1, \ldots, dx_n \; \langle T_f, D\varphi\rangle$$

est également une distribution que l'on note DT_f. Nous justifierons cette notation dans le paragraphe suivant.

(c) Une distribution bien connue des physiciens est la distribution de Dirac, définie par $\langle \delta, \varphi\rangle = \varphi(\vec{0})$. En un point $\vec{a}$ de $\mathbb{R}^n$, la distribution de Dirac $\delta(\vec{a})$ est définie par $\langle \delta(\vec{a}), \varphi\rangle = \varphi(\vec{a})$.

Distributions mathématiques et distributions des charges en physique

Considérons un solide δ de volume V dans $\mathbb{R}^n$ et de densité $f(\vec{x})$. $f(\vec{x})$ représente la masse d'un élément de volume dV contenant le point $\vec{x}$. Appelons T_f la distribution définie par la fonction f [*cf. exemple* **(a)**]. La masse totale du solide S est donnée par l'intégrale :

$$\int_{V \subset \mathbb{R}^n} f(\vec{x})\, dx_1, \ldots, dx_n = \langle T_f, 1\rangle.$$

De même, le moment d'inertie par rapport à l'origine est égal à : $\int_V f(\vec{x}) \| \vec{x} \|^2 dx_1, \ldots, dx_n$, soit $\langle T_f, \|\cdot\|^2\rangle$ où $\|\cdot\|^2$ désigne la fonction $x \mapsto \|x\|^2$. Ainsi donc, en physique et en mécanique, on est conduit à calculer de nombreuses expressions de la forme $\langle T, \varphi\rangle$ où T représente une « distribution » de masses. Mais c'est en électrostatique, en magnétisme et en mécanique quantique que l'on est conduit à faire un usage plus spécifique des distributions (dans l'exemple pris à l'instant, la notion de « mesure » suffit !). Considérons par exemple la distribution de charges électriques définie par un doublet de moment électrique + 1 placé en O sur la droite $\mathbb{R}$. Ce doublet est « la limite » du système T_ε de deux charges $\left(+\frac{1}{\varepsilon} \text{ et } -\frac{1}{\varepsilon}\right)$ placées aux points 0 et ε.

$$\underset{\left(+\frac{1}{\varepsilon}\right)}{\overset{0}{+}} \text{———————————} \underset{\left(-\frac{1}{\varepsilon}\right)}{\overset{\delta}{+}}$$

T_ε définit une *distribution* dont la valeur en une fonction φ de $\mathfrak{D}$ est donnée par $\langle T_\varepsilon, \varphi\rangle = \frac{1}{\varepsilon}\varphi(\varepsilon) - \frac{1}{\varepsilon}(\varphi(0))$. Lorsque $\varepsilon \to 0$, $\langle T_\varepsilon, \varphi\rangle$ tend vers $\varphi'(0)$. Nous sommes ainsi amenés naturellement à définir le doublet par la distribution $\langle T, \varphi\rangle = \varphi'(0)$.

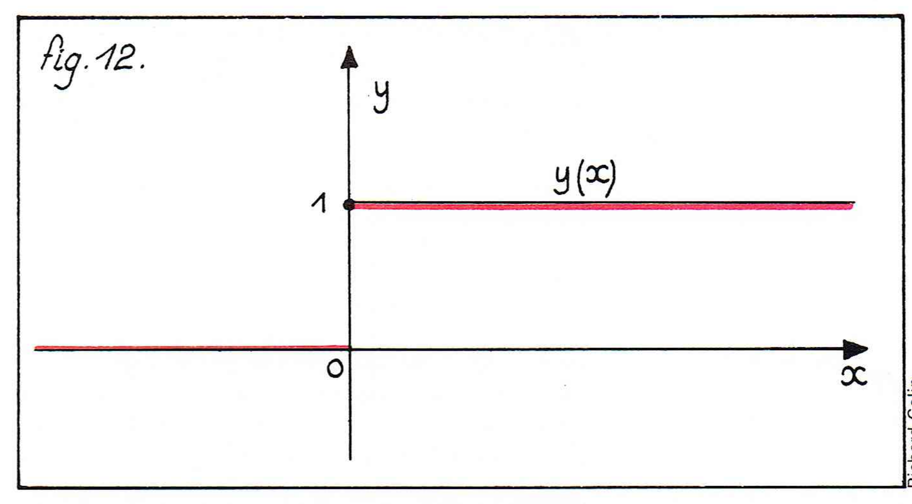
Dérivation des distributions

Cherchons à définir la dérivée $\frac{\partial T}{\partial x_1}$ d'une distribution T sur $\mathbb{R}^n$ par rapport à la variable x_1, de telle sorte que, lorsque T est une fonction f continue et à dérivées partielles premières continues, on retrouve $\frac{\partial f}{\partial x_1}$ dans le sens usuel du calcul différentiel classique. Soit donc f une telle fonction :

$$\left\langle \frac{\partial f}{\partial x_1}, \varphi \right\rangle = \int_{\mathbb{R}^n} \frac{\partial f}{\partial x_1} \cdot \varphi\, dx_1, \ldots, dx_n =$$

$$\int_{x_2} \cdots \int_{x_n} dx_2, \ldots, dx_n \int_{-\infty}^{+\infty} \frac{\partial f}{\partial x_1} \cdot \varphi\, dx_1$$

en intégrant par parties l'intégrale simple $\int_{-\infty}^{+\infty} \frac{\partial f}{\partial x_1} \cdot \varphi\, dx_1$, on trouve $\left[f \cdot \varphi\right]_{-\infty}^{+\infty} - \int_{-\infty}^{+\infty} f \cdot \frac{\partial \varphi}{\partial x_1} dx_1$. La fonction φ étant à support compact, le terme $\left[f \cdot \varphi\right]_{-\infty}^{+\infty}$ disparaît. Finalement, on obtient l'égalité $\left\langle \frac{\partial f}{\partial x_1}, \varphi \right\rangle = -\left\langle f, \frac{\partial \varphi}{\partial x_1}\right\rangle$.



▲ ***Figure 12 : graphe de la fonction d'Heaviside.***

Cette égalité a été obtenue dans le cas où f était différentiable. Nous sommes conduits à la prendre *comme définition* de la dérivée $\frac{\partial T}{\partial x_1}$ en posant :

$$\left\langle \frac{\partial T}{\partial x_1}, \varphi \right\rangle = -\left\langle T, \frac{\partial \varphi}{\partial x_1}\right\rangle$$

On vérifie que cette égalité définit bien $\frac{\partial T}{\partial x_1}$ comme une distribution : il s'agit évidemment d'une forme linéaire sur $\mathfrak{D}$ dont on montre qu'elle est continue pour la topologie définie sur $\mathfrak{D}$.

Plus généralement, on définira les dérivées de tous ordres d'une distribution T par la formule :

$$\langle D^p T, \varphi\rangle = (-1)^{|p|} \langle T, D^p \varphi\rangle$$

où $p = (p_1, \ldots, p_n)$ est un système de n nombres entiers avec $|p| = p_1 + p_2 + \ldots + p_n$, et D^p représente la dérivation $\left(\frac{\partial}{\partial x_1}\right)^{p1} \left(\frac{\partial}{\partial x_2}\right)^{p2} \ldots \left(\frac{\partial}{\partial x_n}\right)^{pn}$.

L'intérêt de cette définition est tout à fait primordial : dans le cas où la distribution T est une fonction intégrable (pas forcément continue), il est possible de définir les dérivées de tous ordres de la fonction f mais il ne s'agit pas en général de fonctions, il s'agit de *distributions*. Ainsi se trouvent justifiées de manière rigoureuse les pratiques des physiciens en calcul symbolique lorsque ceux-ci « dérivaient » des fonctions discontinues.

Exemple : la fonction d'Heaviside (*fig.* 12).

Soit y la fonction de $\mathbb{R} \to \mathbb{R}$ définie par $y(x) = \begin{cases} 0 & x < 0 \\ 1 & x \geqslant 0 \end{cases}$

Cette fonction est utilisée en électricité sous le nom « échelon unité ». Elle n'est bien sûr pas continue en 0. Calculons la dérivée DY :

$$\langle DY, \varphi\rangle = -\langle Y, D\varphi\rangle =$$

$$-\int_{-\infty}^{+\infty} y(x) \cdot \varphi'(x)\, dx = -\int_0^{\infty} \varphi'(x)\, dx =$$

$$-\left[\varphi(x)\right]_0^{+\infty} = \varphi(0) = \langle \delta, \varphi\rangle$$

on a donc DY = δ. δ est la distribution de Dirac [*cf. exemple* **(c)**] que les électriciens appellent « impulsion unité ». La dérivée seconde $D^2 y$ est donnée par

$$\langle D^2 y, \varphi\rangle = \langle D\delta, \varphi\rangle = -\langle \delta, D\varphi\rangle = -\varphi'(0)$$

$D^2 y$ est donc la distribution associée au doublet de moment — 1 en 0.

BIBLIOGRAPHIE

LIONS J.-L., *Cours d'analyse numérique*, Hermann. - ROCKAFELLAR R. T., *Conlex Analysis*, Princeton University. - SCHWARTZ L., *Topologie générale et analyse fonctionnelle*, Hermann, 1970. - *Méthodes mathématiques pour les sciences physiques*, Hermann.

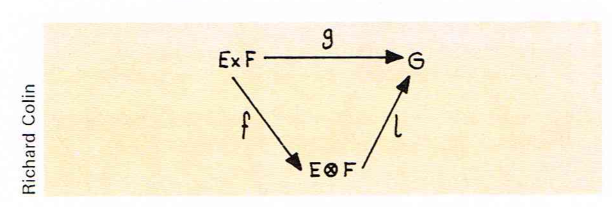
CALCUL TENSORIEL

Le calcul tensoriel provient de problèmes posés par la mécanique classique. C'est à Curbastro Gregorio Ricci (1853-1925) et à Tullio Levi-Civita (1873-1941) que l'on doit le premier exposé relatif à cet outil mathématique abstrait, que, d'un point de vue mathématique, on doit considérer comme une extension de la théorie des espaces vectoriels.

La théorie de la relativité générale, qui a vu le jour grâce à l'existence du calcul, a permis, par ricochet, l'extension de ce domaine au travers des problèmes de géométrie différentielle particulièrement étudiés par Élie Cartan (1869-1951).

La présentation mathématique du calcul tensoriel est en fait celle de l'algèbre multilinéaire ; toutefois, le produit tensoriel de deux espaces vectoriels E et F peut être présenté comme solution — unique à un isomorphisme près — d'un problème dit problème universel : $E \otimes F$ et l'application f, bilinéaire, dite *produit tensoriel*, vérifient la propriété : quels que soient l'espace vectoriel G et l'application bilinéaire $g : E \times F \to G$, il existe $l \in \mathcal{L}(E \otimes F, G)$ telle que $g = l \circ f$.

On dit parfois encore que l'espace vectoriel $E \otimes F$ est l'espace vectoriel des combinaisons linéaires formelles des éléments de $E \times F$; cette application sera sans doute plus clarifiée par la présentation adoptée ci-dessous :



Richard Colin

Produit tensoriel

Considérons un espace vectoriel E, de dimension finie n, sur un corps K. On sait que l'ensemble des applications linéaires de E dans K forme, pour les opérations usuelles d'addition et de produit par un scalaire, un espace vectoriel, appelé dual de E, noté E^*. De plus, si $(e_i)_{i=1,\ldots,n}$ est une base de E, alors les applications e^j définies pour tout $j = 1, 2, \ldots, n$ par :

$$e^j(e_i) = \begin{cases} 0 & \text{si } i \neq j \\ 1 & \text{si } i = j \end{cases}$$

forment une base de E^*.

Considérons alors $f \in E^*$ et $g \in E^*$, deux formes linéaires. L'application $t : E \times E \to K$, définie par

$$t(x, y) = f(x) \cdot g(y),$$

est bilinéaire et représente une forme d'association de f et g par l'intermédiaire des valeurs de chacune *prises séparément*. Cette forme bilinéaire sur E est appelée *produit tensoriel* de f et g ; on la note $f \otimes g$. Plus généralement encore, si $f_1, f_2, \ldots, f_p$ sont p formes linéaires sur E, l'application $T : E^p \to K$ définie par :

$$T(x_1, x_2, \ldots, x_p) = f_1(x_1) \cdot f_2(x_2) \ldots f_p(x_p)$$

est p-linéaire et « associe » $f_1, \ldots, f_p$ à travers les valeurs de chacune ; on l'appelle produit tensoriel de $f_1, f_2, \ldots, f_p$, noté $f_1 \otimes f_2 \otimes f_3 \ldots \otimes f_p$.

En particulier, les éléments $(e^j)_{j=1,\ldots,n}$ de la base de E^* combinés deux à deux par le produit tensoriel définissent n^2 formes bilinéaires $(e^i \otimes e^j)_{\substack{i=1,\ldots,n \\ j=1,\ldots,n}}$ telles que $e^i \otimes e^j(x, y) = x_i y_j$, en désignant par x_i la i-ième coordonnée de x et par y_j la j-ième de y dans la base choisie de E. L'espace vectoriel engendré (c'est-à-dire reconstitué par les opérations d'addition et de produit par un scalaire) par ces formes bilinéaires est désigné par le nom de produit tensoriel de E^* par lui-même, ce qu'on note $E^* \otimes E^*$, ou encore $\bigotimes_2 E$.

Cet espace n'est pas vraiment inconnu ; c'est en réalité l'ensemble des formes bilinéaires sur E. En effet, soit B une telle forme, alors,

$$\text{si } x = \sum_{i=1}^{n} x_i e_i \quad \text{et } y = \sum_{j=1}^{n} y_j e_j$$

sont deux vecteurs quelconques de E, on a :

$$B(x, y) = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i y_j B(e_i, e_j)$$

soit, en posant $\lambda_{i,j} = B(e_i, e_j)$, on voit que :

$$B(x, y) = \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_{i,j} e^i \otimes e^j(x, y)$$

donc que $B \in \bigotimes_2 E$. Tout élément de $E^* \otimes E^*$ étant une forme bilinéaire, l'ensemble de celles-ci est donc exactement $E^* \otimes E^*$.

On définit donc de même l'ensemble

$$\underbrace{E^* \otimes E^* \otimes \ldots \otimes E^*}_{p \text{ fois}} = \bigotimes_p E,$$

qui n'est autre que l'espace vectoriel des formes p-linéaires sur E ; il est engendré par les applications de E^p dans K, $e^{i_1} \otimes e^{i_2} \otimes \ldots \otimes e^{i_p}$ où $i_1, i_2, \ldots, i_p$ désignent p valeurs, non nécessairement distinctes, des nombres entiers $1, 2, \ldots, n$. On peut montrer que ces applications sont linéairement indépendantes pour tous les choix différents — il y en a n^p — des indices $i_1, i_2, \ldots, i_p$; par conséquent, $\dim \bigotimes_p E = n^p$.

Le produit tensoriel ainsi défini tant pour des formes linéaires sur un même espace que pour des espaces tous égaux entre eux peut être généralisé à des espaces différents, mais sur un même corps de base. Prenant deux espaces vectoriels E_1 et E_2, de dimensions n_1 et n_2, sur un corps K et deux formes linéaires $f_1 \in E_1^*$, $f_2 \in E_2^*$ on définit le produit tensoriel $f_1 \otimes f_2$ par :

$$f_1 \otimes f_2(x_1, x_2) = f_1(x_1) \times f_2(x_2), \quad \forall x_1 \in E_1 \quad \text{et} \quad \forall x_2 \in E_2$$

Ce produit tensoriel est une application bilinéaire de $E \times F$ dans K. En appliquant cette opération aux éléments d'une base $(e^i)_{i=1,\ldots,n_1}$ de E^*_1 et d'une base $(e'^j)_{j=1,\ldots,n_2}$ de E^*_2, on définit $n_1 \times n_2$ formes bilinéaires $(e^i \otimes e'^j)_{\substack{i=1,\ldots,n_1 \\ j=1,\ldots,n_2}}$ qui engendrent un espace vectoriel, dit produit tensoriel de E^* et F^*, noté $E^* \otimes F^*$, qui n'est autre que l'ensemble des formes bilinéaires sur $E \times F$ (ce que l'on montre par un raisonnement analogue à celui fait pour $F = E$).

L'opération *produit tensoriel* apparaît donc comme un moyen de représentation de la multilinéarité par des outils linéaires, ou en sens inverse comme une « recomposition », par de simples formes linéaires, de formes multilinéaires. Elle permet de dégager un cadre général pour l'algèbre multilinéaire.

On peut noter dès maintenant que la possibilité d'identifier à un isomorphisme près le dual de E^* (soit le bidual E^{**}) à l'ensemble E permet de construire le produit tensoriel $\underbrace{E \otimes E \otimes \ldots \otimes E}_{p \text{ fois}}$, noté aussi $\overset{p}{\otimes} E^*$, ainsi que le produit tensoriel $x_1 \otimes x_2 \otimes \ldots \otimes x_p$ qui est une application p-linéaire de E^{*p} dans K définie par :

$$x_1 \otimes x_2 \otimes \ldots \otimes x_p(f_1, f_2, \ldots, f_p) = f_1(x_1) \times f_2(x_2) \times \ldots \times f_p(x_p)$$

quelles que soient $f_1 \in E^*, f_2 \in E^*, \ldots, f_p \in E^*$.

Plus généralement encore, on peut construire le produit tensoriel d'éléments de E et d'éléments de E^*. Soit par exemple deux formes linéaires $f \in E^*$ et $g \in E^*$, et trois éléments X_1, X_2, X_3 de E (donc formes linéaires sur E^*) ; le produit tensoriel $f \otimes g \otimes X_1 \otimes X_2 \otimes X_3$ est une forme 5-linéaire :

$$(x_1, x_2, h_1, h_2, h_3) \to f(x_1) \times g(x_2) \times h_1(X_1) \times h_2(X_2) \times h_3(X_3)$$

définie sur $E^2 \times E^{*3}$. On peut donc former les n^5 applications

$$e^{i_1} \otimes e^{i_2} \otimes e_{i_3} \otimes e_{i_4} \otimes e_{i_5} \quad \text{de } E^2 \times E^{*3} \text{ dans K}$$

où i_1, i_2, i_3, i_4, i_5 désignent cinq valeurs, non nécessairement distinctes, des nombres $1, 2, \ldots, n$. On note

$$E \otimes E \otimes E^* \otimes E^* \otimes E^*$$

l'espace vectoriel engendré par ces n^5 formes linéaires.

C'est par l'idée du produit tensoriel d'espaces que l'on peut maintenant dégager la notion de tenseur et une loi de composition, dite produit tensoriel, entre ces êtres mathématiques abstraits.

Tenseurs

Un élément de l'espace dual E* est dit *tenseur* 1-*covariant* (ou tenseur covariant d'ordre 1). Une forme p-linéaire sur E est dite *tenseur p-covariant* (ou tenseur covariant d'ordre p).

La dénomination choisie s'explique ainsi : si u est un élément de E*, on peut écrire $u = \sum_{i=1}^{n} u_i \cdot e^i$ dans la base $(e^i)_{i=1,\dots,n}$ duale de celle $(e_i)_{i=1,\dots,n}$ choisie pour E, avec $u_i = u\,(e_i)$; si $(\varepsilon_i)_{i=1,\dots,n}$ désigne une autre base de E, liée à la première par les formules de passage $\varepsilon_i = \sum_{j=1}^{n} \alpha_{ji} e_j$, on peut écrire dans la base $(\varepsilon^i)_{i=1,\dots,n}$ duale de $(\varepsilon_i)_{i=1,\dots,n}$ une décomposition $u = \sum_{i=1}^{n} v_i \varepsilon^i$ où $v_i = u\,(\varepsilon_i)$. Alors v_i s'écrit aussi :

$v_i = u\left(\sum_{j=1}^{n} \alpha_{ji} e_j\right) = \sum_{j=1}^{n} \alpha_{ji} u_j$. Les formules de passage pour les coordonnées (ou composantes) des éléments de E* et pour une base de E sont donc les mêmes : on dit pour cela que l'on a *covariance*.

Un élément de l'espace E (pouvant être, puisqu'on suppose E de dimension finie, considéré comme une forme linéaire sur E*) est dit *tenseur* 1-*contravariant* (ou tenseur contravariant d'ordre 1). Une forme q-linéaire définie sur E*, c'est-à-dire une application de $(E^*)^q$ dans le corps de base K, linéaire par rapport à chaque argument, est appelée *tenseur q-contravariant* (ou tenseur contravariant d'ordre q). En effet, il y a variation opposée des formules de passage pour les coordonnées d'un vecteur rapporté à une base et les vecteurs de cette base puisque, si $\varepsilon_i = \sum_{j=1}^{n} \alpha_{ji} e_j$, alors $(x^i)_{i=1,\dots,n}$ et $(y^j)_{j=1,\dots,n}$ coordonnées de x respectivement selon les bases $(e_i)_{i=1,\dots,n}$ et $(\varepsilon_j)_{j=1,\dots,n}$ sont liées par : $x^i = \sum_{j=1}^{n} \alpha_{ij} y^j$.

L'algèbre tensorielle fait, on le voit, un usage intensif d'indices. Cela ne saurait surprendre celui qui reconnaît dans cette partie des mathématiques une présentation des phénomènes de multilinéarité. Toutefois, l'usage en devient vite lourd et la répétition des symboles de sommations concernant les divers indices en jeu apparaît dès lors comme un fardeau d'écriture (et de lecture !) bien pesant.

Un jeu de conventions d'écritures permet de tout simplifier. Tout d'abord, on affecte à tout indice qui a une signification de covariance ou de contravariance une position en rapport : les indices de covariance seront supérieurs (sans risque de confusion avec un exposant) tandis que les indices de contravariance seront placés en dessous. Ainsi, on note

$$x = \sum_{i=1}^{i=n} x_i e^i \quad \text{si } x \in E^*$$

alors que

$$y = \sum_{i=1}^{n} y^i e_i \quad \text{si } y \in E\,;$$

par suite, on notera $\varepsilon_i = \sum_{i=1}^{n} \alpha_i{}^j e_j$ les formules de changement de base dans E.

Dans ces formules de sommation, on remarque que l'indice par rapport auquel s'effectue la somme se trouve toujours une fois en position basse et une fois en position haute ; ceci n'est rien d'autre que la transcription de l'idée de dualité (ou de correspondance par dualité). Cette constatation amène à poser la convention dite *convention d'Einstein* :

Dans tout monôme où figure un même indice répété une fois comme indice covariant (position supérieure) et une fois comme indice contravariant (position inférieure), on considère qu'il s'agit en fait de la somme des monômes obtenus pour les diverses valeurs de cet indice (on sous-entend donc la sommation). Par conséquent, avec cette convention, on écrira $x = x_i e^i$ au lieu de $x = \sum_{i=1}^{n} x_i e^i$, ou bien $\varepsilon_i = \alpha^j{}_i e_j$ au lieu de $\varepsilon_i = \sum_{j=1}^{n} \alpha_i{}^j e_j$, ou bien encore : $x^k{}_p{}^l = \alpha^i{}_h{}^j \cdot a_i{}^k \cdot b_j{}^l \cdot c_p{}^h$ au lieu de l'expression usuelle $x^k{}_p{}^l = \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{h=1}^{q} \alpha^i{}_h{}^j \cdot a_i{}^k \cdot b^{jl} \cdot c^p{}_h$.

Deux remarques s'imposent au sujet de l'utilisation de cette règle. Tout d'abord, le domaine des valeurs d'un indice de sommation n'est plus précisé, et il importe donc bien qu'il n'y ait aucune ambiguïté, ce qui demande une pratique du problème d'autant plus grande que le nombre des indices en jeu est élevé. En second lieu, il est clair que la position d'un indice est primordiale, car la convention d'Einstein ne peut s'appliquer que si l'on a la possibilité d'associer covariance et contravariance. Par conséquent, chaque indice doit être placé en conformité avec sa signification : covariant ou contravariant. On est donc tenté de voir une difficulté supplémentaire dans l'application de ce mécanisme ; ce qui, en réalité est inexact, car l'utilisation simultanée d'indices, toujours délicate, demande en outre la connaissance de leurs domaines de valeurs et de leur signification.

Il a été montré que l'ensemble des tenseurs p-covariants munis des deux opérations usuelles — addition et produit par un scalaire — avait une structure d'espace vectoriel ; on le désigne par $\bigotimes_p E$. Il en est de même pour l'ensemble des tenseurs q-contravariants, $\bigotimes^q E$. Une base de l'espace $\bigotimes_p E$ est obtenue par les éléments

$$e^{i_1} \otimes e^{i_2} \otimes \dots \otimes e^{i_p},$$

produit tensoriel de p vecteurs de la base duale de celle choisie dans E, et une base de l'espace $\bigotimes^q E$ est obtenue par les éléments $e_{j1} \otimes e_{j2} \otimes \dots \otimes e_{jq}$, produit tensoriel de q vecteurs de la base de E (identifiée à la base du bidual E**). Pour tout élément de $\bigotimes_p E$ ou de $\bigotimes^q E$, il existe donc un système de n^p scalaires dans le premier cas, de n^q dans le second, réalisant la décomposition relativement à la base citée : ce sont — conformément à la terminologie usuelle de l'algèbre linéaire — les coordonnées du tenseur relativement à la base donnée.

Ainsi, soit T un tenseur 3-covariant, $T \in \bigotimes_3 E$; ce tenseur est une application de E^3 dans K, et pour tout triplet

$(x, y, z) \in E^3$ où $x = x^i e_i$, $y = y^j e_j$, $z = z^k e_k$ (convention d'Einstein), on a :

$$T\,(x, y, z) = T\,(x^i e_i, y^j e_j, z^k e_k) = x^i y^j z^k\, T\,(e_i, e_j, e_k)$$

en développant selon la trilinéarité ; en posant

$$T\,(e_i, e_j, e_k) = t_{ijk},$$

on obtient $T\,(x, y, z) = t_{ijk} x^i y^j z^k$, ce qui donne

$$T = t_{ijk} e^i \otimes e^j \otimes e^k$$

par définition des éléments $e^i \otimes e^j \otimes e^k$.

De la même façon, pour un tenseur 4-contravariant $U \in \bigotimes^4 E$, on obtient $U = u^{ijkl} e_i \otimes e_j \otimes e_k \otimes e_l$ en posant $u^{ijkl} = U\,(e^i, e^j, e^k, e^l)$.

L'obtention, et donc l'utilisation, des coordonnées d'un tenseur est identique à celle de la matrice d'une application linéaire. On verra plus loin qu'il ne s'agit pas d'une coïncidence.

Les tenseurs mixtes se définissent selon le même principe. Par exemple, un tenseur 2-covariant, 1-contravariant, 1-covariant est une forme 4-linéaire sur l'espace $E^2 \times E^* \times E$; l'ensemble de ces tenseurs forme un espace vectoriel qui est engendré par les éléments

$$e^i \otimes e^j \otimes e_k \otimes e^l,$$

les indices i, j, k et l prenant toutes les valeurs entières de

1 à $n = \dim E$. Si T est un tenseur de ce type, on a :

$$T(x, y, F, z) = T(x^i e_i, y^j e_j, F_k e^k, z^l e_l)$$

d'où $T = T_{ij}{}^k{}_l \, e^i \otimes e^j \otimes e_k \otimes e^l$

avec $T_{ij}{}^k{}_l = T(e_i, e_j, e^k, e_l)$.

Les n^4 scalaires $T_{ij}{}^k{}_l$ sont les coordonnées du tenseur mixte T.

Pour tout tenseur, covariant, contravariant ou mixte, les coordonnées sont liées au choix de la base choisie pour l'espace E, puisque alors la base duale (celle de E*) est définie, et donc celle de tout produit tensoriel d'espaces. Dans tout problème, le choix de la base peut nécessiter une modification ultérieure et ne pas être irrévocable. Il est donc nécessaire de disposer — en algèbre tensorielle comme en algèbre linéaire — de formules établissant l'effet d'un changement de base sur un tenseur, donc sur ses coordonnées. Dans le cas d'un tenseur T covariant d'ordre 3, si l'on passe de la base $(e_i)_{i=1,\ldots,n}$ à la base $(\varepsilon_i)_{i=1,\ldots,n}$ selon $\varepsilon_i = \alpha^j{}_i e_j$ (convention d'Einstein), on a :

$T = t_{ijk} \, e^i \otimes e^j \otimes e^k$ dans la première base, et

$T = v_{ijk} \, \varepsilon^i \otimes \varepsilon^j \otimes \varepsilon^k$ dans la seconde ; donc

$$v_{ijk} = T(\varepsilon_i, \varepsilon_j, \varepsilon_k) = T(\alpha^l{}_i \, e_l, \alpha^m{}_j \, e_m, \alpha^r{}_k \, e_r)$$

soit

$$v_{ijk} = \alpha^l{}_i \, \alpha^m{}_j \, \alpha^r{}_k \, T(e_l, e_m, e_r) = \alpha^l{}_i \, \alpha^m{}_j \, \alpha^r{}_k \, t_{lmr}.$$

On établirait aussi bien pour un tenseur 4-contravariant U ayant les coordonnées u^{ijkl} dans la base (e_i) et w^{ijkl} dans la base (ε_i) :

$$w^{ijkl} = \beta^i{}_m \, \beta^j{}_p \, \beta^k{}_r \, \beta^l{}_s \, u^{mprs}$$

où $\beta^i{}_m$ représente l'élément situé à la i-ième ligne et m-ième colonne dans la matrice inverse de celle formée par les $\alpha^j{}_i$.

Pour un tenseur mixte T, 1-covariant, 1-contravariant, 1-covariant de coordonnées $t_i{}^j{}_k$ et $v_i{}^j{}_k$, on obtiendrait la formule :

$$v_i{}^j{}_k = \alpha^l{}_i \, \beta^j{}_m \, \alpha^r{}_k \, t_l{}^m{}_r.$$

En général, on voit que les coefficients α interviennent pour chaque covariance et les coefficients β pour chaque contravariance. Bien entendu, on pourrait écrire une formule générale de changement de base, mais son intérêt serait très limité en raison de sa complexité d'écriture, ce qu'il est aisé d'imaginer.

Il est à noter que ces expressions réalisent une condition nécessaire et suffisante pour qu'un système de scalaires puisse être considéré comme les coordonnées d'un tenseur de type défini. De cette règle, d'ailleurs, dérivent encore trois autres critères de reconnaissance d'un tenseur par ses coordonnées.

Quelques tenseurs particuliers

A toute application linéaire $L \in \mathcal{L}(E)$, on associe un tenseur mixte T du second ordre 1-contravariant, 1-covariant en posant :

$$T : (F, x) \to F(L(x)) \text{ où } F \in E^* \text{ et } x \in E.$$

Les coordonnées de ce tenseur sont données par :

$$t^i{}_j = T(e^i, e_j) = e^i(L(e_j))\,;$$

or $L(e_j)$ est un élément de E dont les coordonnées — selon la base $(e_i)_{i=1,\ldots,n}$ — forment la j-ième colonne de la matrice de L dans la base choisie, et donc sa j-ième composante — soit $e^i(L(e_j))$ — n'est autre que l'élément $m^i{}_j$ situé à la i-ième ligne et j-ième colonne de cette matrice. On en déduit donc l'existence d'une bijection entre l'ensemble $\mathcal{L}(E)$ des applications linéaires de E dans lui-même et l'ensemble $E^* \otimes E$ des tenseurs mixtes 1-contravariant, 1-covariant, puisque les formules usuelles de changement de base montrent que les éléments d'une matrice peuvent être considérés comme les coordonnées d'un tel tenseur mixte.

Si L est une application linéaire de E* dans E, on peut lui associer un tenseur 2-covariant $T \in E^* \otimes E^*$ par :

$$T : (F, G) \to F(L(G)) \quad \text{où } F \in E^* \quad \text{et } G \in E^*.$$

Les coordonnées de ce tenseur sont :

$$t^{ij} = T(e^i, e^j) = e^i(L(e^j))\,;$$

c'est-à-dire que, comme précédemment, les coordonnées de T sont les éléments de la matrice de l'application L. On met donc ainsi en évidence une bijection entre l'ensemble $\underset{2}{\otimes} E$ (soit $E^* \otimes E^*$) et l'ensemble $\mathcal{L}(E^*, E)$.

On peut noter, pour préciser ces liens entre ensembles de tenseurs d'ordre 2 et ensembles d'applications linéaires (ou aussi ensemble des matrices carrées d'ordre n), que ce sont des bijections linéaires, c'est-à-dire des isomorphismes.

Le tenseur — dit *tenseur de Kronecker* — mixte du second ordre 1-covariant, 1-contravariant défini par ses composantes :

$$\delta^i{}_j = 0 \quad \text{si } i \neq j, \quad \delta^i{}_j = 1 \quad \text{si } i = j,$$

correspond — par l'isomorphisme vu plus haut — à l'application identité de l'espace E. On sait que la matrice de cette application est invariante par changement de la base de E ; cette propriété reste vraie pour les composantes du tenseur de Kronecker. En effet, si les coordonnées dans une autre base sont $t^i{}_j$, on peut écrire : $t^i{}_j = \beta^i{}_l \, \alpha^m{}_j \, \delta^l{}_m$, ce qui devient, après avoir simplifié les termes nuls de la somme sous-entendue : $t^i{}_j = \beta^i{}_j \, \alpha^l{}_j$, soit $t^i{}_j = \delta^i{}_j$ puisque les coefficients α et β sont ceux de deux matrices inverses l'une de l'autre.

Parmi les tenseurs du second ordre 2-covariants ou 2-contravariants, on distingue deux cas particuliers : les tenseurs *symétriques* et les tenseurs *antisymétriques*. Dans le premier cas, il s'agit de tenseurs T tels que

$$T(X, Y) = T(Y, X)$$

pour tous $(X, Y) \in E \times E$ pour le cas covariant, pour tous $(X, Y) \in E^* \times E^*$ pour la contravariance ; dans le second, il s'agit de tenseurs U tels que

$$U(X, Y) = -U(Y, X).$$

Il est clair qu'on ne peut étendre ces notions aux tenseurs mixtes. Un tenseur 2-covariant sera donc symétrique si ses coordonnées vérifient $t_{ij} = t_{ji}$ pour tout couple (i, j) ; pour un tenseur 2-contravariant symétrique, on aura $t^{ij} = t^{ji}$. Un tenseur 2-covariant sera antisymétrique si ses coordonnées vérifient $t_{ij} = -t_{ji}$ pour tout couple (i, j) ; pour un tenseur 2-contravariant antisymétrique, on a $t^{ij} = -t^{ji}$.

Tout tenseur 2-covariant (de même que tout tenseur 2-contravariant) peut s'écrire comme somme d'un tenseur symétrique et d'un tenseur antisymétrique de mêmes espèces, puisque, si T est de coordonnées t_{ij}, alors le tenseur de composante $\frac{1}{2}(t_{ij} + t_{ji})$ est symétrique, le tenseur de composantes $\frac{1}{2}(t_{ij} - t_{ji})$ est antisymétrique, et de plus :

$$t_{ij} = \frac{1}{2}(t_{ij} + t_{ji}) + \frac{1}{2}(t_{ij} - t_{ji}).$$

Mais bien évidemment, d'après les isomorphismes vus plus haut, les propriétés des tenseurs d'ordre 2 se recoupent avec celles des applications linéaires.

Un tenseur p-covariant est dit alterné s'il s'annule lorsqu'on l'applique à p vecteurs dont deux sont égaux. On voit que, si $p = 2$, un tel tenseur est antisymétrique. On définit de même les tenseurs q-contravariants alternés. Il est alors facile de montrer qu'un tenseur est alterné dès que, dans une base fixée, ses coordonnées sont des fonctions alternées des indices. Par conséquent, un tenseur alterné d'ordre $p > \dim E = n$ est nul, puisque, pour toute coordonnée $t_{i_1 i_2 \ldots i_p}$, les indices $i_1, i_2, \ldots, i_p$ prennent leurs valeurs parmi $1, 2, \ldots, n$, donc deux d'entre eux au moins sont égaux.

Lorsque l'espace vectoriel E est un espace euclidien (c'est-à-dire muni d'une forme bilinéaire symétrique dont la forme quadratique associée est définie positive), il est possible d'associer les tenseurs de même ordre des types définis plus haut (on parle de *tenseurs affines*) : on parle alors de *tenseur euclidien*. Si l'on effectue le produit scalaire de deux éléments X et Y de E, on écrit

$$\langle X, Y \rangle = \langle x^i e_i, y^j e_j \rangle = x^i y^j \langle e_i, e_j \rangle$$

(convention d'Einstein) ; soit $\langle X, Y \rangle = g_{ij} x^i y^j$, si l'on pose $g_{ij} = \langle e_i, e_j \rangle$ (les signes $\langle\ \rangle$ désignent ici le résultat de la forme bilinéaire symétrique donnant à E son caractère euclidien). En tenant compte de ce que le produit scalaire est invariant par changement de base, et qu'il est

commutatif (car la forme bilinéaire est supposée symétrique), les scalaires g_{ij} sont les composantes (covariantes) d'un tenseur euclidien, que l'on appelle le *tenseur fondamental de l'espace E*.

Ceci n'est valable que parce que tout espace euclidien est isomorphe à son dual — donc peut lui être identifié — par l'intermédiaire de la correspondance qui, à tout $x \in E$, associe la forme linéaire définie par $y \rightarrow \langle x, y \rangle$ pour tout $y \in E$. Les propriétés de la forme bilinéaire fondamentale (ici désignée par $\langle \, \rangle$) font de cette application un isomorphisme d'espaces vectoriels qui ne dépend aucunement d'une quelconque base de E : c'est donc bien un isomorphisme canonique qui permet de passer de E à E* ou, si l'on préfère, de remplacer, lorsque besoin est, un élément $x \in E$ par un élément de E* sans qu'aucune propriété linéaire en soit affectée.

Ainsi, un vecteur (contravariant) $x \in E$ se décompose sur une base $(e_i)_{i=1,...,n}$ selon $x = x^i e_i$, les scalaires x^i étant appelés composantes contravariantes de x. L'élément $f_x \in E^*$ associé à x par l'isomorphisme précédent admet pour composantes, notées $(f_x)_i$, dans la base duale $(e^i)_{i=1,...,n}$ $(f_x)_i = \langle x, e_i \rangle = g_{ij}x^j$. Puisque l'on peut identifier x et f_x, il est donc légitime de donner à ces scalaires, qui sont les composantes covariantes d'un élément de E*, le nom de composantes covariantes de x ; on les note donc x_i, et elles vérifient $x_i = g_{ij}x^j$ (sommation sous-entendue).

Par conséquent, ainsi que le montre la construction des tenseurs affines (éléments d'un produit tensoriel d'espaces E ou E*), de tels tenseurs, d'un même ordre, définissent un seul et même tenseur dit *tenseur euclidien* dont les différentes composantes covariantes, contravariantes ou mixtes se déduisent entre elles par produit (et sommation sous-entendue) par g_{ij} ou g^{ij} (éléments de la matrice inverse) ; ces opérations étant *au plus* répétées N fois, si N désigne l'ordre du tenseur (c'est-à-dire la somme du degré de covariance et du degré de contravariance).

Opérations sur les tenseurs

C'est par les opérations de calcul tensoriel que l'on va dégager la puissance de cet outil très abstrait ; les structures obtenues montrent sa valeur par rapport à ce qui a été obtenu dans le cadre linéaire.

Tout d'abord, concernant les tenseurs d'un même type (c'est-à-dire de même degré de covariance et contravariance), on construit la structure usuelle d'espace vectoriel par l'intermédiaire d'une loi additive et d'un produit par un scalaire. L'addition d'un tenseur T et d'un tenseur T' de même espèce est donc définie par un tenseur T + T' tel que :

$$(T + T')\,(X_1, X_2, ..., X_N) = T\,(X_1, X_2, ..., X_N) + T'\,(X_1, X_2, ..., X_N)$$

ce qui est possible puisque T et T' opèrent sur les mêmes espaces. Il est clair que cette addition donne à l'ensemble des tenseurs d'un même type une structure de groupe commutatif. Le produit d'un tenseur T par un scalaire α est un tenseur du même type, αT, tel que :

$$(\alpha T)\,(X_1, X_2, ..., X_N) = \alpha \cdot T\,(X_1, X_2, ..., X_N).$$

Cette loi externe complète alors la structure linéaire de l'ensemble des tenseurs d'une même espèce : p-covariant ou q-contravariant, ou bien mixte m-covariant et n-contravariant.

Mais, dépassant cette structure, on dispose d'une opération supplémentaire, le produit tensoriel, qui va permettre de définir sur de vastes ensembles de tenseurs un cadre d'algèbre.

Si T_1 est un tenseur $\binom{p}{q}$ — p-covariant, q-contravariant — et T_2 un tenseur d'espèce $\binom{m}{n}$, on définit le produit tensoriel $T_1 \otimes T_2$ comme un tenseur d'espèce $\binom{p+m}{q+n}$, tel que :

$$T_1 \otimes T_2\,(X_1, ..., X_p, Y_1, ..., Y_q, X'_1, ..., X'_m, Y'_1, ..., Y'_n) = T_1\,(X_1, ..., X_p, Y_1, ..., Y_q) \cdot T_2\,(X'_1, ..., X'_m, Y'_1, ..., Y'_n).$$

Le premier tenseur est en fait une forme $(p + q)$-linéaire, le second une forme $(m + n)$-linéaire ; leur produit tensoriel est une forme $(p + m + q + n)$-linéaire. Il ne s'agit donc pas d'une loi de composition interne sur l'ensemble des tenseurs d'un type donné.

Les coordonnées du produit tensoriel s'obtiennent par produit des coordonnées concernées des tenseurs dont on fait le produit. Ainsi, si T_1 est un tenseur $\binom{3}{2}$ et T_2 un tenseur $\binom{2}{1}$, le produit tensoriel $T_1 \otimes T_2$ est d'espèce $\binom{5}{3}$ et ses coordonnées sont

$$T_1 \otimes T_2\,(e_i, e_j, e_k, e^h, e^l, e_r, e_s, e^t) = T_1\,(e_i, e_j, e_k, e^h, e^l) \times T_2\,(e_r, e_s, e^t)$$

donc $(T_1 \otimes T_2)_{i,j,k,}{}^{h,l,}{}_{r,s,}{}^{t} = (T_1)_{i,j,k,}{}^{h,l,} \cdot (T_2)_{r,s,}{}^{t}$.

Les principales propriétés du produit tensoriel sont données par :

$$U \otimes (T_1 + T_2) = U \otimes T_1 + U \otimes T_2 ;$$
$$(T_1 + T_2) \otimes U = T_1 \otimes U + T_2 \otimes U \text{ (distributivité)}$$
$$T \otimes (\lambda \cdot U) = (\lambda T) \otimes U = \lambda \cdot (T \otimes U).$$
$$(T_1 \otimes T_2) \otimes T_3 = T_1 \otimes (T_2 \otimes T_3) \text{ (associativité).}$$

Par conséquent, l'application : $(T_1, T_2) \rightarrow T_1 \otimes T_2$ est bilinéaire. On peut noter que si T_1 et T_2 sont deux tenseurs, leur produit de composition au sens des applications — lorsqu'il a un sens — n'est pas en général égal à leur produit tensoriel : $T_1 \circ T_2 \neq T_1 \otimes T_2$.

Le produit d'un tenseur p-covariant par un tenseur m-covariant est un tenseur $(p + m)$-covariant. Il vient donc à l'idée d'associer les espaces $\underset{p}{\otimes} E$ de façon à regrouper tous les tenseurs covariants. Pour cela, on définit l'ensemble $\underset{\omega}{\otimes} E$ comme somme directe des espaces vectoriels $\underset{p}{\otimes} E$, donc $\underset{\omega}{\otimes} E = \overset{\infty}{\underset{p=0}{\oplus}} \left(\underset{p}{\otimes} E\right)$ avec :

$$\underset{o}{\otimes} E = E, \quad \underset{1}{\otimes} E = E^*, \quad \underset{2}{\otimes} E = E^* \otimes E^*, \text{ etc.}$$

Tout élément de $\underset{\omega}{\otimes} E$, dit tenseur affine covariant, sera donc une somme de tenseurs p-covariants ($p \in \mathbb{N}$) dont seul un nombre fini est non nul. Il devient alors possible de définir une addition et un produit tensoriel à l'intérieur de $\underset{\omega}{\otimes} E$. Les différentes propriétés que l'on établit alors aisément munissent cet ensemble d'une structure d'anneau qui, combinée à celle d'espace vectoriel, en fait *l'algèbre des tenseurs covariants sur E*.

Cette construction est évidemment réalisable pour les tenseurs contravariants, et *l'algèbre des tenseurs contravariants sur E*, $\overset{\omega}{\otimes} E$, est définie par :

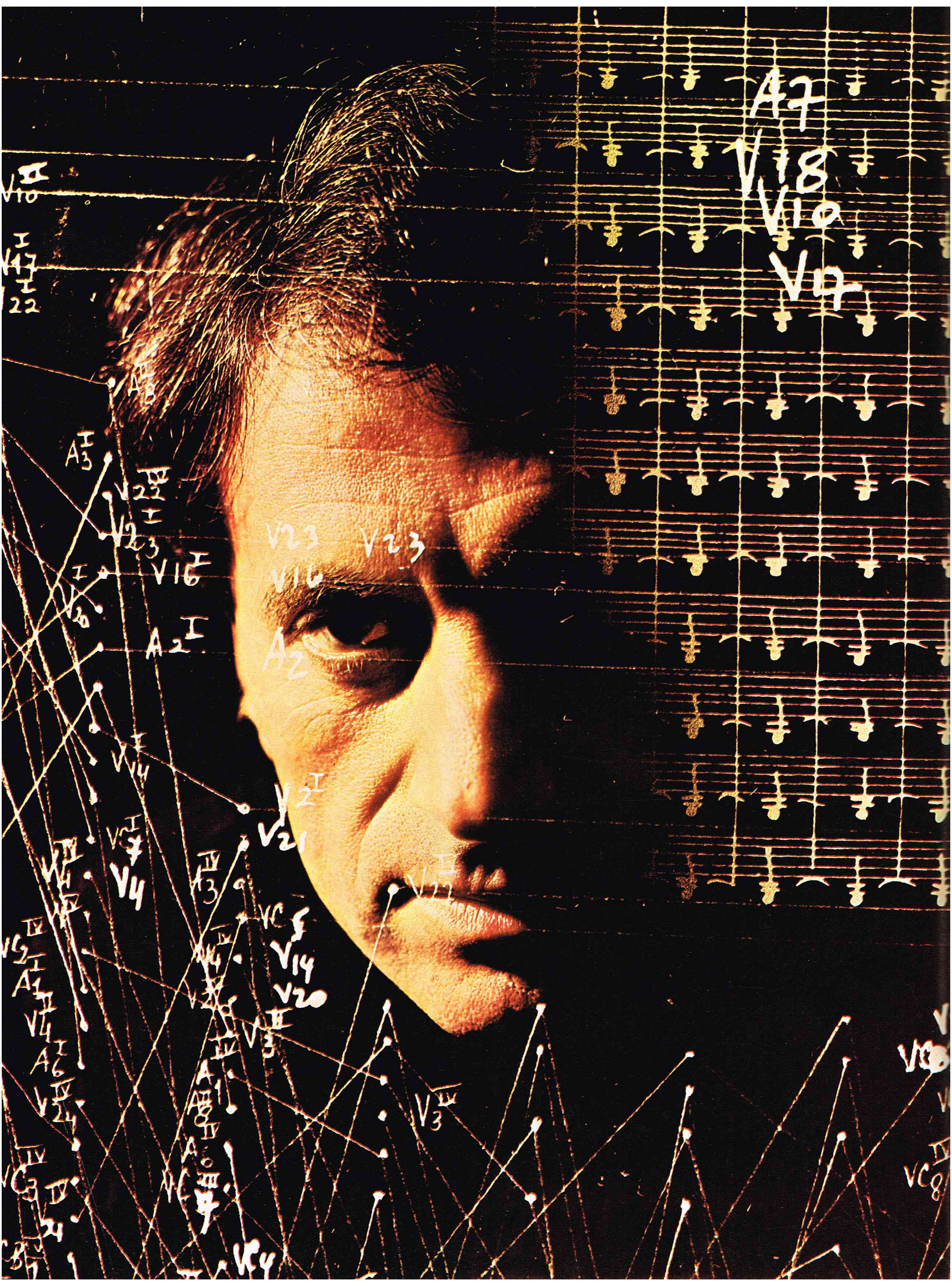
$$\overset{\omega}{\otimes} E = \overset{\infty}{\underset{q=0}{\oplus}} \left(\overset{q}{\otimes} E\right),$$

de façon analogue, avec $\overset{0}{\otimes} E = K$ (corps de base),

$$\overset{1}{\otimes} E = E, \quad \overset{2}{\otimes} E = E \otimes E, \text{ etc.}$$

En dehors des opérations fondamentales, addition et produit tensoriel, définies plus haut, il existe encore une opération qui, à un tenseur d'ordre N, associe un tenseur N-2, il s'agit de la *contraction*, qui consiste à égaler un indice de covariance et un indice de contravariance (pour faire donc après la somme sous-entendue par la convention d'Einstein), ce qui fait donc jouer aux espaces E et E* leur rôle de dualité ; d'un tenseur de type $\binom{p}{q}$, on obtient donc un tenseur d'ordre $\binom{p-1}{q-1}$. Cette opération pouvant être répétée tant que le tenseur obtenu est mixte. Par exemple, on sait identifier une matrice carrée à un tenseur mixte $\binom{1}{1}$ du second ordre dont les coordonnées $\alpha_j{}^i$ sont les coefficients de la matrice. La contraction lui associe un tenseur d'ordre 0, donc un scalaire qui n'est autre que la trace de la matrice $\Sigma\, \alpha_{ii}$.

L'utilisation combinée de cette opération et du produit tensoriel définit alors un *produit contracté*, obtenu par contraction d'indices dans le produit tensoriel. Ainsi, en prenant T défini par les composantes $T^{ij}{}_k$ et U par les composantes $U^l{}_m$, on contracte le produit $T \otimes U$ de composantes $T^{ij}{}_k\, U^l{}_m = V^{ij}{}_k{}^l{}_m$ en égalant j et k, ce qui donne un tenseur contracté $(T \otimes U)_c$ de composantes $V^{ij}{}_j{}^l{}_m$, soit $W^{il}{}_m$, que l'on peut à nouveau contracter pour


A7
V18
V10
V19

obtenir un tenseur de composantes $W^{im}{}_m$, soit Z^i après sommation.

Deux exemples montrent bien qu'il s'agit d'un cadre abstrait général pour représenter des opérations déjà connues, mais dont la nature différente ne permettait pas de les associer.

Prenons $T \in E^*$ défini par ses coordonnées t_i et $X \in E$ défini par ses coordonnées x^i ; le produit $T \otimes X$ est défini par ses coordonnées $t_i x^j$ que l'on note $u_i{}^j$, et le produit contracté $(T \otimes X)_c$ est alors le scalaire $u_i{}^i = t_i x^i$, qui n'est autre que le produit scalaire usuel $\langle T, X \rangle$.

En prenant maintenant un tenseur T mixte du second ordre, associé à une application linéaire L de matrice $(\alpha_j{}^i)$ et $X \in E$ défini par ses coordonnées x^k, le produit $T \otimes X$ a pour composantes $u_j{}^{ik} = \alpha_j{}^i x^k$; la contraction $j = k$ donne un tenseur de composantes $y^i = \alpha^i{}_j x^j$, c'est-à-dire que $(T \otimes X)_c = L(X)$.

Les tenseurs et les opérations définies sur eux permettent donc de définir une idée générale regroupant celles déjà obtenues par la linéarité, puis la multilinéarité.

Enfin, une autre opération établit un lien avec l'analyse et la géométrie différentielle moderne : *le produit extérieur*.

Dans l'ensemble $\overset{2}{\otimes} E$ des tenseurs 2-covariants sur E, les expressions $e^i \otimes e^j - e^j \otimes e^i$, que l'on note $e^i \wedge e^j$, sont des tenseurs alternés. La famille de ces vecteurs forme une base du sous-espace vectoriel A_2 des tenseurs alternés 2-covariants $\left(\text{sous-espace de } \underset{2}{\otimes} E\right)$, qui est donc de dimension $\dfrac{n(n-1)}{2}$.

Plus généralement, dans l'ensemble $\underset{p}{\otimes} E$, les expressions

$$e^{i_1} \wedge e^{i_2} \wedge \ldots \wedge e^{i_p} = \sum_{\sigma \in S_p} \varepsilon(\sigma)\, e^{i\sigma(1)} \otimes e^{i\sigma(2)} \otimes \ldots \otimes e^{i\sigma(p)}$$

(où σ désigne une permutation du groupe « symétrique » S_p, de signature $\varepsilon(\sigma)$ [c'est-à-dire $\varepsilon(\sigma) = (-1)^N$ où N désigne le nombre de transpositions de σ], pour

$$1 \leqslant i_1 < i_2 < \ldots < i_p \leqslant n)$$

sont des tenseurs alternés p-covariants. La famille de ces tenseurs forme une base du sous-espace vectoriel A_p des tenseurs alternés p-covariants (sous-espace de $\underset{p}{\oplus} E$) qui est donc de dimension $\dfrac{n!}{p!\,(n-p)!}$. Cet espace vectoriel peut être ainsi considéré comme le produit extérieur de E p fois par lui-même, ce que l'on note $\underset{p}{\wedge} E = A_p$.

De l'idée ainsi développée pour E et ses vecteurs de base, on dérive l'idée généralisée de produit extérieur de deux tenseurs, application bilinéaire de $A_p \times A_q$ dans A_{p+q} correspondant au produit tensoriel, application bilinéaire de $\underset{p}{\otimes} E \times \underset{q}{\otimes} E$ dans $\underset{p+q}{\otimes} E$; mais ici le produit tensoriel est insuffisant, car deux tenseurs alternés n'ont pas en général un produit tensoriel alterné.

On définit pour tout tenseur p-covariant T un tenseur p-covariant alterné $\overline{T}$, dit *antisymétrisé* de T, par :

$$\overline{T}(x_1, x_2, \ldots, x_p) = \frac{1}{p!} \sum_{\sigma \in S_p} \varepsilon(\sigma)\, T(x_{\sigma(1)}, x_{\sigma(2)}, \ldots, x_{\sigma(p)})$$

en conservant les notations vues ci-dessus.

Par exemple,

$$\text{si } p = 2, \overline{T}(x_1, x_2) = \frac{1}{2}\left[T(x_1, x_2) - T(x_2, x_1)\right].$$

Lorsque T est lui-même alterné, alors $\overline{T} = T$; cette condition est donc nécessaire et suffisante pour qu'un tenseur soit alterné. Par conséquent, l'opération d'antisymétrisation applique $\underset{p}{\otimes} E$ dans son sous-espace $\underset{p}{\wedge} E$; c'est en fait une projection.

Le produit extérieur de deux tenseurs T et T' est alors défini comme l'antisymétrisé de leur produit tensoriel, soit : $T \wedge T' = \overline{T \otimes T'}$; et cette opération est bien une application bilinéaire de $A_p \times A_q$ dans A_{p+q}, qui est associative et non commutative ; on a, en fait,

$$T' \wedge T = (-1)^{pq}\, T \wedge T'.$$

De même que l'on a construit $\underset{\omega}{\otimes} E$ pour disposer d'une structure respectant les produits tensoriels, on construit *l'algèbre extérieure des tenseurs alternés covariants* :

$$\underset{\omega}{\wedge} E = \bigoplus_{p=0}^{\infty} A_p, \qquad \text{en posant } A_0 = K.$$

Bien évidemment, $\underset{\omega}{\wedge} E$ est une sous-algèbre de $\underset{\omega}{\otimes} E$.

Les principes de dualité qui guident tout ce qui précède s'appliquent encore et permettent de définir, de manière duale, les produits extérieurs $e_{i_1} \wedge e_{i_2} \wedge e_{i_3} \wedge \ldots \wedge e_{i_q}$, et donc de poser $A^{[q]} = \overset{q}{\wedge} E$, sous-espace vectoriel de $\overset{q}{\otimes} E$, de dimension $\dfrac{n!}{q!\,(n-q)!}$, ainsi que l'antisymétrisation comme application projetant $\overset{q}{\otimes} E$ sur $\overset{q}{\wedge} E$. Le produit extérieur de deux tenseurs contravariants définit alors une application de $A^{[m]} \times A^{[q]}$ dans $A^{[m+q]}$, qui conduit à poser l'algèbre extérieure des tenseurs alternés contravariants $\overset{\omega}{\wedge} E$ comme la somme directe $\bigoplus_{p=0}^{\infty} A^{[p]}$, sous-algèbre de $\underset{\omega}{\otimes} E$. Dans ce cas, on peut donc définir le produit extérieur de m vecteurs de E, analogue au produit extérieur de p formes linéaires sur E, puisque l'on peut identifier (par isomorphisme) tout vecteur de E à un élément du dual E^*.

Les produits extérieurs permettent de tracer un cadre simple pour les déterminants, en se plaçant dans l'espace $A^{[n]}$ où $n = \dim E$. Cet espace de dimension 1 admet la base $e^1 \wedge e^2 \wedge \ldots \wedge e^n$; la valeur de cette forme pour le n-uplet de vecteurs $(X_1, X_2, \ldots, X_n)$ est le déterminant de ces n vecteurs par rapport à la base $(e_1, e_2, \ldots, e_n)$. Le calcul de déterminants trouve ainsi sa place normale dans le cadre général de l'algèbre multilinéaire qu'est l'algèbre tensorielle.

L'algèbre extérieure trouve aussi une issue particulièrement féconde dans le calcul différentiel sur une variété.

Parmi les nombreuses applications des tenseurs, la théorie de la relativité utilise les notions fondamentales d'analyse tensorielle sur les espaces euclidiens réels et sur les espaces riemanniens.

Pour tout vecteur X d'un espace euclidien réel E muni d'une base (e_i), les composantes contravariantes x^i sont définies par $X = x^i e_i$. La différentiation de cette égalité selon les règles classiques donne :

$$dX = dx^i e_i + x^i de_i.$$

Chaque vecteur de_i peut être répété par ses composantes contravariantes que l'on note $\omega^j{}_i$, c'est-à-dire que :

$$de_i = \omega^j{}_i e_j.$$

Par conséquent, après avoir redistribué les indices, on peut écrire :

$$dX = (dx^i + x^k \omega^i{}_k)\, e_i,$$

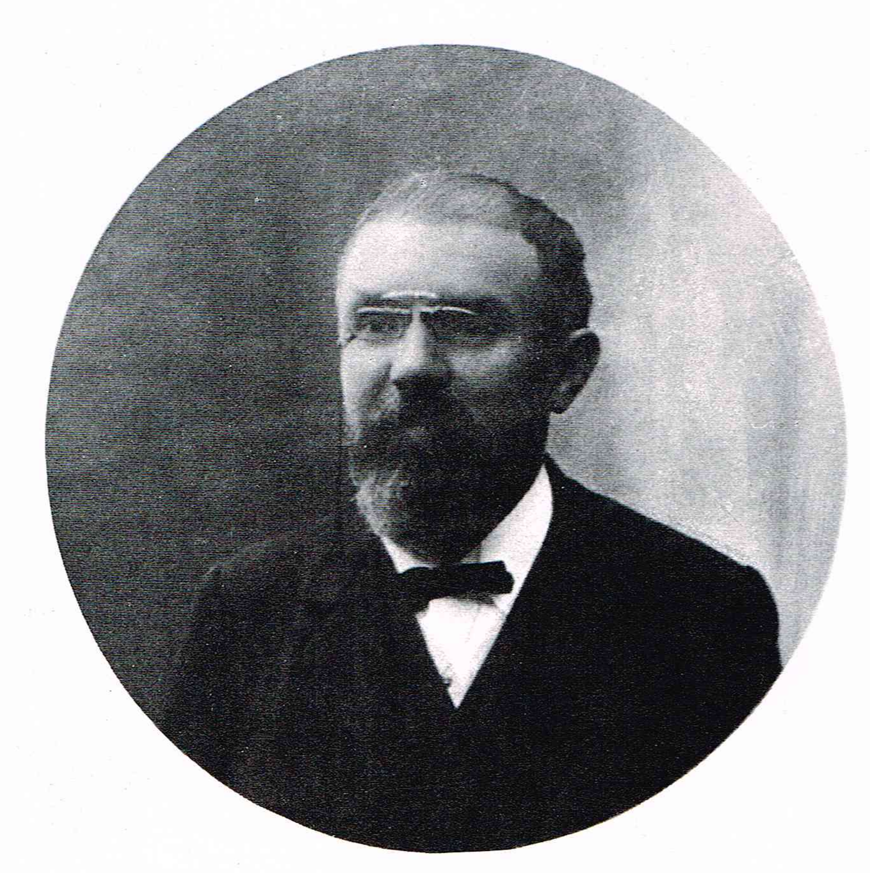
ce qui montre que les composantes contravariantes du vecteur dX sont les quantités $dx^i + \omega^i{}_k x^k$ que l'on nomme *différentielles absolues de* x^i pour chaque valeur de i.

Cette démarche peut être répétée, de façon analogue, pour un tenseur euclidien. Un résultat important, *le théorème de Ricci*, a trait au tenseur fondamental g_{ij} : la différentielle absolue du tenseur fondamental est nulle.

L'utilisation des coordonnées curvilignes permet d'autre part de poser la notion de dérivée covariante, et de définir les symboles de Christoffel de première et deuxième espèces dont l'utilisation est très fréquente pour le calcul différentiel classique sur les champs de vecteurs.

BIBLIOGRAPHIE

BOURBAKI M., *Algèbre*, ch. I à III, Hermann, 1971. - CARTAN É., *Leçons sur la géométrie des espaces de Riemann*, Gauthier-Villars, 1946. - DELACHET A., *le Calcul tensoriel*, P. U. F., 1969, coll. « Que sais-je ? ». - GOUYON R., *Calcul tensoriel*, Vuibert, 1963. - LEVI-CIVITA T., *Lezioni di calcolo differenziale assoluto*, 1925. - LICHNEROWICZ A., *Algèbre et analyse linéaires*, Masson, 1947 ; *Éléments de calcul tensoriel*, A. Colin, 1950. - SCHWARTZ L., *les Tenseurs*, Hermann, 1975.



Palais de la Découverte, Paris

▲ *Le mathématicien français* **Henri Poincaré** *(1854-1912).*

COURBES ET SURFACES

L'*Analysis situs*, maintenant nommée topologie, a pris un second essor dès la fin du XIXe siècle. La topologie algébrique s'est alors développée, prenant naissance essentiellement dans les travaux du mathématicien français *H. Poincaré* (1854-1912) pour devenir une branche autonome en pleine expansion grâce aux nouvelles techniques algébriques de classification.

La dimension

Une notion très utilisée est celle de *dimension*. Il est courant de dire que la droite est de dimension 1, le plan de dimension 2, l'espace ordinaire de dimension 3 et plus généralement que $\mathbb{R}^n$ est de dimension n. Essayons de trouver une définition générale. La notion de dimension ne peut pas appartenir purement et simplement à la théorie des ensembles puisqu'on peut établir une correspondance univoque et continue, mais non biunivoque entre $\mathbb{R}$ et $\mathbb{R}^n$. La dimension devra donc être définie par l'intermédiaire de considérations de topologie.

Une première approche peut être la suivante : un plan peut être recouvert de rectangles (on pense alors à une paroi recouverte de bandes [*fig.* 1-*a*]). On peut trouver des points appartenant à *trois* rectangles, mais on peut éviter (en recourant à un recouvrement approprié) la présence de points appartenant à quatre rectangles. Une région spatiale peut être recouverte avec des parallélépipèdes : on peut toujours trouver des points communs à *quatre* parallélépipèdes et éviter la présence de points communs à plus de quatre parallélépipèdes (*fig.* 1-*b*). Le nombre critique, écrit en italique *(trois, quatre)*, dépasse donc de 1 celui qu'on appelle naturellement la dimension de l'espace.

Une approche assez naturelle peut être faite à l'aide de l'observation suivante : le contour d'une région tridimensionnelle est bidimensionnel, celui d'une région bidimensionnelle est unidimensionnel : en réalité, les choses ne sont pas toujours ainsi; la définition d'un contour doit alors être élaborée de façon plus précise, et c'est l'objet de l'*homologie*.

Si E est un espace métrisable, on dit que celui-ci est :

(1) de dimension — 1 s'il est vide;

(2) de dimension $\leqslant n$ ($n = 0, 1, 2, ...$) en un de ses points x lorsque, pour tout voisinage V de x, il existe un voisinage U de x avec $U \subseteq V$, dont le contour est de dimension $\leqslant n - 1$;

(3) de dimension $\leqslant n$ si en tout point sa dimension est $\leqslant n$;

(4) de dimension n en un de ses points x lorsque sa dimension en x est $\leqslant n$, mais non $\leqslant n - 1$;

(5) de dimension n si sa dimension est $\leqslant n$, mais non $\leqslant n - 1$.

Pour $n = 0$, **(2)** indique qu'un espace est de dimension 0 en x si tout voisinage de x en contient un autre dont le contour est vide; si cela arrive pour tout $x \in E$, la dimension de E est 0. On peut ensuite obtenir les espaces de dimension 1, et ainsi de suite.

Par exemple, un espace qui contient un nombre fini de points est de dimension 0, $\mathbb{R}^n$ est de dimension n, tandis que $\mathbb{Q}^n$ est de dimension 0.

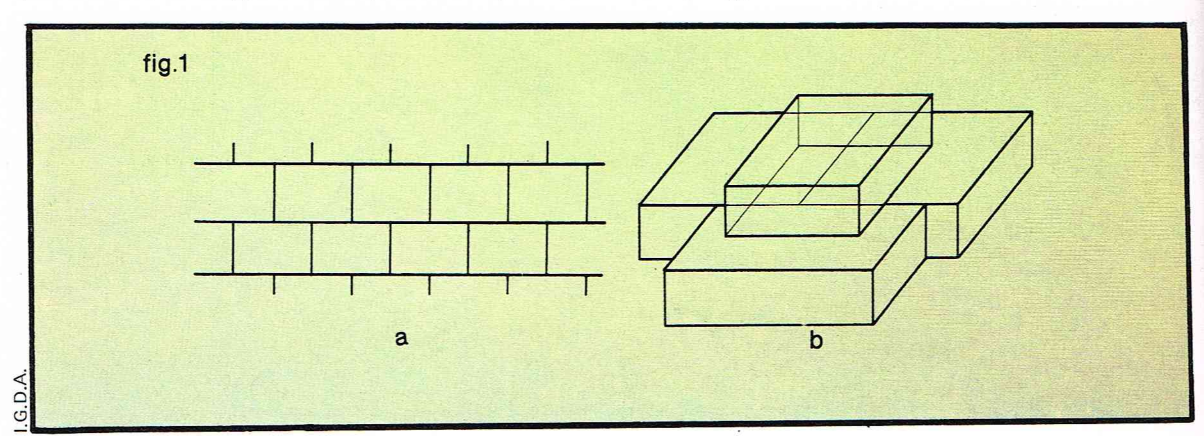
Notons la dimension de E par dim E, on a :

$$\dim (A \cup B) \leqslant \dim A + \dim B + 1 ;$$
$$\dim (A \times B) \leqslant \dim A + \dim B$$

si A et B ne sont pas tous les deux vides.

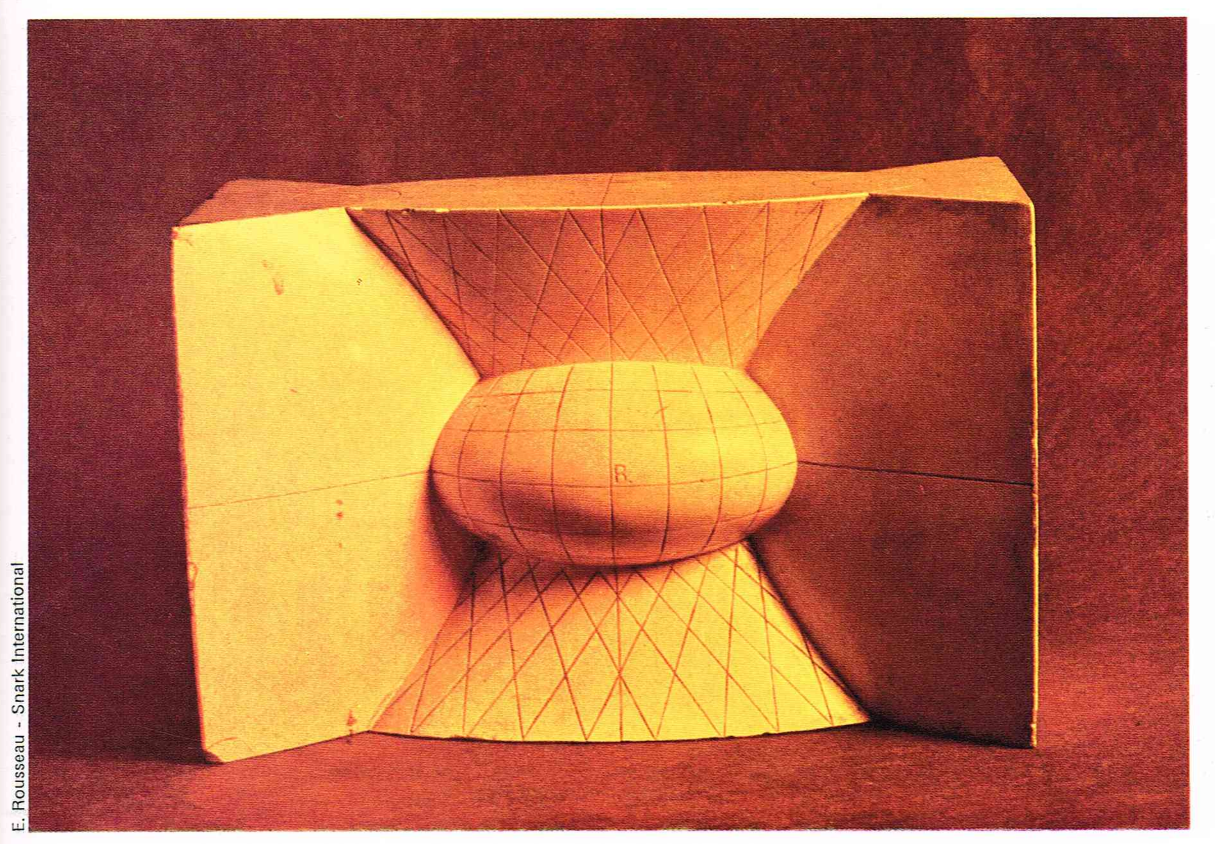
La notion de dimension permet de définir d'une façon générale celle de courbe. Les courbes planes peuvent être définies comme des ensembles fermés et connexes (un sous-espace topologique E' de E est connexe s'il n'existe pas deux ouverts disjoints de E, U_1 et U_2, tels que $E' \subset U_1 \cup U_2$) ne contenant aucun point intérieur; mais on ne peut pas définir d'une façon analogue les courbes non planes. Plus généralement, une courbe est un ensemble fermé et connexe à une dimension, cette définition étant topologiquement invariante. On définit l'*ordre* d'un point p d'une courbe comme le plus petit entier n tel qu'il y ait des voisinages de p, de diamètre arbitrairement petit, et dont les frontières contiennent au plus n points de la courbe.

La notion de courbe continue est importante; on appelle ainsi le *chemin* décrit par un point qui se déplace continûment. On peut donc dire qu'une courbe continue est une image par une application univoque et continue du segment [0, 1]. Le même point peut être l'image de plusieurs points du segment, et dans ce cas, il est point multiple de la courbe, ainsi la *courbe de Peano* (carré considéré comme une courbe de $\mathbb{R}^2$ et image du segment [0, 1]).



I.G.D.A.

▶ ***Figure 1a-1b :*** *exemple graphique simplifié d'un procédé intuitif pour parvenir à la définition topologique de la notion de dimension.*



◀ *Toute surface peut être triangulée, c'est-à-dire qu'on peut la diviser en un nombre fini ou infini dénombrable de triangles. Ici, une surface courbe.*

▼ *Ci-dessous, figure 2 :* **triangulation** *d'une sphère (1) ; d'un cylindre (2) ; d'un ruban de Möbius (3) ; d'un tore (4).*
En bas, figure 3 : exécutant les identifications indiquées, on obtient la **bouteille de Klein** *(à gauche) et le* **plan projectif** *(à droite).*

Triangulation d'une surface

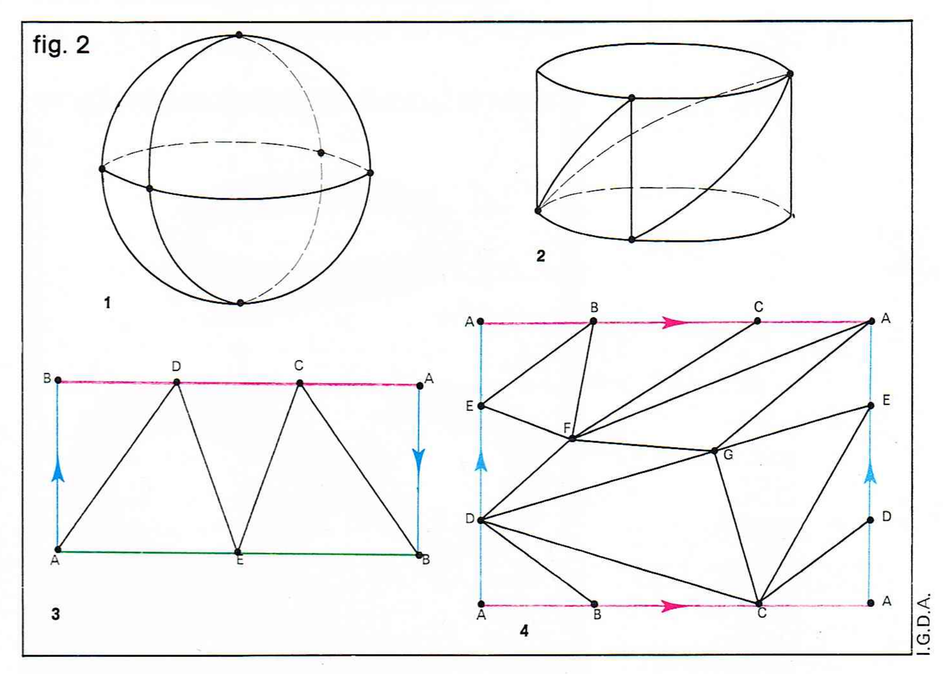
Les faces d'un polyèdre de $\mathbb{R}^3$ sont des polygones, et on peut diviser un polygone en des triangles (voir *Géométrie*). La surface d'un polyèdre peut ainsi être considérée comme un système de triangles qui vérifie les conditions :

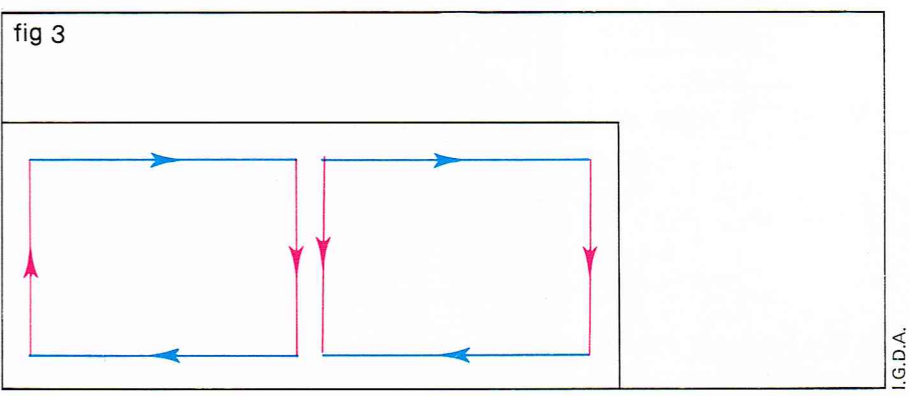
(1) un point intérieur à un triangle n'appartient à aucun autre triangle ;
(2) chaque côté appartient à deux triangles qui ne possèdent comme points communs que ceux qui appartiennent à ce côté (triangles adjacents) ;
(3) les triangles qui ont un sommet commun définissent un cycle, deux triangles consécutifs étant adjacents ;
(4) entre deux triangles quelconques, on peut insérer une chaîne de triangles, deux triangles consécutifs étant adjacents.

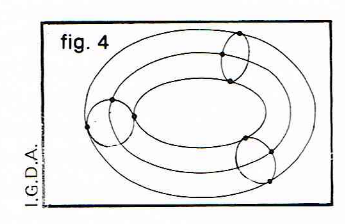
Toute image topologique d'un triangle ordinaire est appelée *triangle* (gauche). Toute surface peut être triangulée, c'est-à-dire qu'on peut la diviser en un nombre fini (si la surface est compacte) ou infini dénombrable de triangles satisfaisant aux conditions qui viennent d'être énoncées. Pour une même surface, il existe plusieurs triangulations vérifiant les quatre conditions mentionnées.

On dit que deux surfaces sont équivalentes si on peut passer de l'une à l'autre par une déformation continue. On peut ainsi classer les surfaces, les surfaces compactes étant les plus faciles à classer. Tout d'abord, on distingue les surfaces orientables et les surfaces non orientables. Toute surface possède une décomposition en triangles (pour les hypersurfaces, il s'agira de simplexes) ; sur chaque triangle, on peut définir un sens de parcours (il en existe deux différents). Si, pour la surface S, chaque triangle de sa décomposition peut être orienté de telle façon que l'ensemble des orientations soit *cohérent*, on dit alors que la surface est orientable (*fig.* 2 et 3).

Deux surfaces compactes sont équivalentes si et seulement si elles ont même type d'orientabilité et même caractéristique d'Euler-Poincaré. La caractéristique d'une surface est un nombre entier qu'on calcule après avoir réalisé sa triangulation (on peut toujours décomposer une surface compacte en un nombre fini de polygones) : si S est le nombre de sommets, A le nombre d'arêtes, F le nombre de faces, on démontre que $S - A + F$ est indépendant de la triangulation ; ainsi dans l'espace à trois dimensions, on a :







▲ *Figure 4 : pour le tore, la caractéristique* S — A + F *est zéro ; dans cette représentation « polyédrique », on a en fait :* S = *9,* A = *18,* F = *9.*

$$S - A + F = 2$$

pour tout polyèdre convexe ; ce nombre est appelé *caractéristique d'Euler-Poincaré* de la surface.

La caractéristique d'une surface orientable s'écrit : $n = 2 - 2\gamma$, avec γ entier positif ou nul, et on appelle γ le *genre* de la surface ; pour une surface non orientable, on a : $n = 2 - \gamma$, avec γ entier supérieur ou égal à 1 (*fig.* 4).

Classer les surfaces consiste à énumérer les différentes classes de surfaces équivalentes. Le genre d'une surface compacte est un outil essentiel et direct pour la classification : ainsi, on montre que pour toute surface compacte orientable de genre γ, il y a équivalence avec un tore à γ poignées (*fig.* 5) [un tore à une poignée étant une surface obtenue en faisant tourner un cercle autour d'une droite de son plan qui ne la coupe pas].

Variété topologique et différentiable

La topologie nous permet de préciser proprement les concepts de *courbe, surface*...

En géométrie euclidienne et en géométrie analytique, on rencontre des types particuliers de courbes et de surfaces (par exemple, les coniques et les quadriques). Nous verrons quelles sont les difficultés qui se présentent lorsqu'on cherche à passer à une définition générale. La droite s'identifie (en géométrie analytique) avec l'ensemble $\mathbb{R}$. On dit qu'une courbe est un ensemble équipotent à $\mathbb{R}$, mais ce concept est trop général puisque le plan est aussi équipotent à $\mathbb{R}$. En entendant par courbe un espace homéomorphe à $\mathbb{R}$, on exclut par exemple la circonférence, et maintenant notre notion est trop restreinte. Une courbe est plutôt un espace *localement* homéomorphe à $\mathbb{R}$, c'est-à-dire homéomorphe morceau par morceau (chaque point possède un voisinage homéomorphe à un intervalle), et une surface est un espace localement homéomorphe à $\mathbb{R}^2$.

Tout ceci amène à la définition de *variété topologique*. Une variété topologique V_n est un espace qu'on peut recouvrir d'ouverts, chacun d'eux étant homéomorphe à une boule ouverte de $\mathbb{R}^n$: la variété qui en résulte est de dimension n. Tout homéomorphisme d'un ouvert V_n sur une boule ouverte de $\mathbb{R}^n$ est appelé une *carte* de V_n. Si $x \in V_n$: il peut arriver que x appartienne au domaine de deux cartes φ et Ψ (*fig.* 6). Alors, en lui associant les points du type $\varphi(x)$, $\Psi(x)$, on obtient un homéomorphisme d'un ouvert de $\mathbb{R}^n$ sur un autre : celui-ci s'appelle *changement de carte*. Si $\varphi(x)$ ainsi que $\Psi(x)$ sont des n-uplets de nombres réels : $\varphi(x) = (x_1, x_2, ..., x_n)$ et $\Psi(x) = (y_1, y_2, ..., y_n)$, alors le changement de carte se représente par les équations :

$$y_i = f_i(x_k)$$

les fonctions f_i étant continues et inversibles.

On dit que V_n est une variété différentiable de *classe* C^r (ou une r-variété) si elle est une variété topologique dotée d'une carte telle que tout changement de carte se passe par des fonctions f_i de classe C^r, c'est-à-dire continues et admettant des dérivées (par rapport aux x_k) continues au moins jusqu'à l'ordre r. On remarquera que les variétés de classe C^s, pour $s > r$, sont des cas particuliers de celles qui sont de classe C^r. S'il existe des dérivées continues de tous ordres, on parle de classe C^∞, et si les f_i sont des fonctions développables en séries entières au voisinage de chaque point, on parle de *variétés analytiques* ou de classe C^ω.

On dit aussi que les cartes φ, Ψ... définissent un *atlas* de classe C^r, de $\mathbb{R}^n$ sur V_n ; l'ensemble des points de tous les atlas de classe C^r, pour lesquels les changements réciproques de carte sont de classe C^r, définissent sur V_n une *structure de variété différentiable*.

Par exemple, une surface sphérique est une variété analytique représentable avec non moins de deux cartes (*fig.* 7) ; un plan projectif est une variété analytique représentable avec non moins de trois cartes ; un tore est aussi une variété analytique, etc.

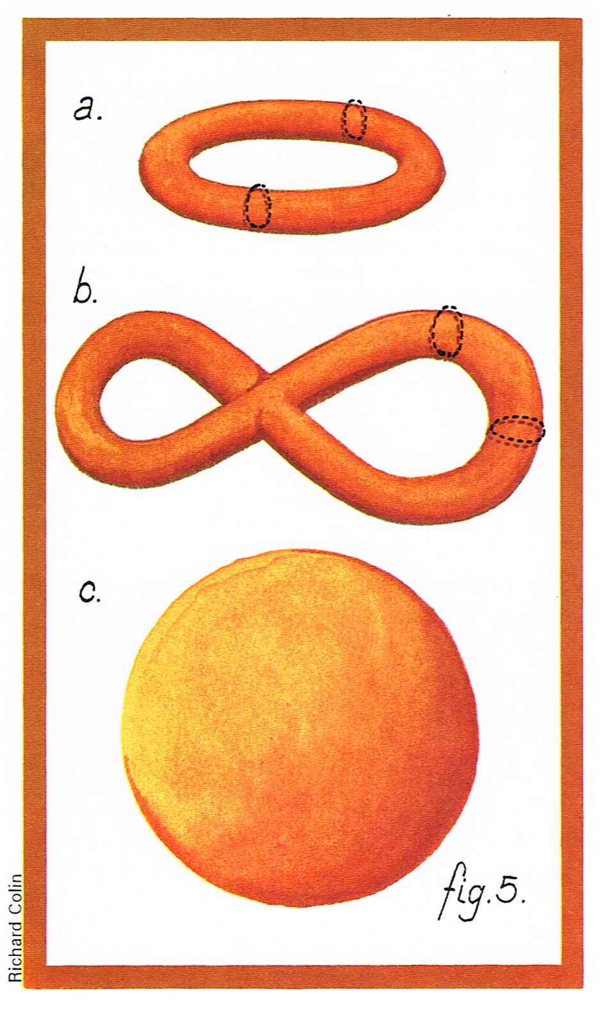
Simplexes

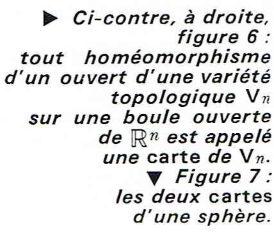
Pour généraliser les notions et les résultats de triangulation, de surface orientée, etc., à des courbes et des surfaces, non plus dans des espaces de dimension trois, mais dans des espaces de dimension quelconque, on a recours à des outils plus abstraits. Les simplexes et la topologie algébrique (homologie, homotopie) sont à la base de cette extension.

Une branche assez importante de la topologie est la *topologie algébrique*, dans laquelle on applique de façon intensive les notions et les techniques de l'algèbre abstraite. Par exemple, la formule d'Euler s'obtient en faisant appel à la topologie algébrique. Plus généralement, pour un polyèdre dont la surface n'est pas homéomorphe à celle d'une sphère, le nombre S — A + F varie : par exemple, il prend la valeur zéro pour le tore (*fig.* 5). L'idée globale est d'exprimer des propriétés topologiques par l'intermédiaire de caractères algébriques (la théorie des groupes y intervient de façon fondamentale) et numériques.

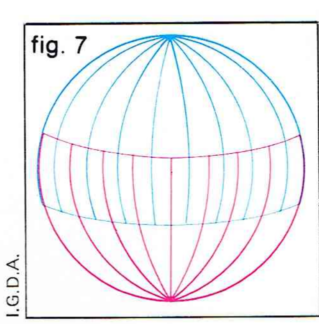
On se place dans un espace euclidien de dimension n, E. Soit $x_0, x_1, ..., x_p$, $p + 1$ points linéairement indépendants dans E. Alors, l'ensemble S des points $(y^1, y^2, ..., y^n)$ de E vérifie :

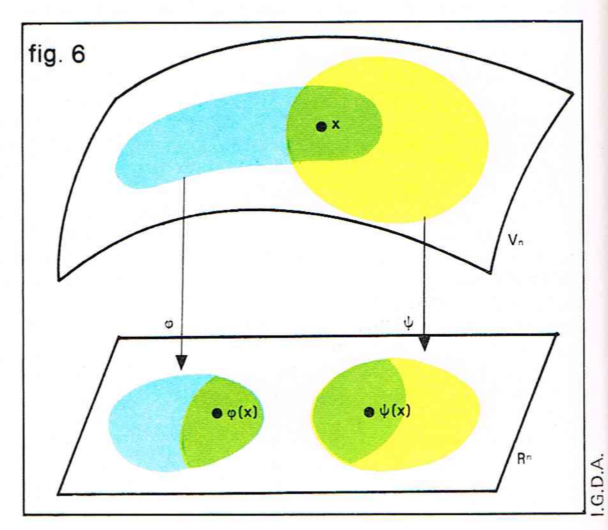
▶ *Figure 5 : a, tore à une poignée :* $\gamma = 1$, $n = 0$; F = *2,* S = *2,* A = *4.* *b, tore à deux poignées :* $\gamma = 2$; $n = -2$. *c, sphère (tore sans poignées) :* $\gamma = 0$, $n = 2$.





▶ *Ci-contre, à droite, figure 6 : tout homéomorphisme d'un ouvert d'une variété topologique* V_n *sur une boule ouverte de* $\mathbb{R}^n$ *est appelé une carte de* V_n.
▼ *Figure 7 : les deux cartes d'une sphère.*





$$y^i = \sum_{k=0}^{k=p} \lambda_k x^i{}_k \qquad (\text{pour } 1 \leqslant i \leqslant n)$$

$$\lambda_k \leqslant 0 \qquad \text{et} \qquad \sum_{k=0}^{k=p} \lambda_k = 1$$

s'appelle le *p-simplexe euclidien de sommets* $x_0, x_1, ..., x_p$, noté $(x_0, x_1, ..., x_p)$.

Par exemple, si $p = 2$, soit z le point qui divise le segment x_1x_2 dans le rapport $\frac{\lambda_2}{\lambda_1}$, alors le point qui divise le segment $x_0 z$ dans le rapport $\frac{\lambda_1 + \lambda_2}{\lambda_0}$ est bien un point du 2-simplexe, qui est donc formé par les points du triangle $x_0x_1x_2$.

On peut raisonner de manière analogue pour montrer que le 3-simplexe de sommets x_0, x_1, x_2, x_3 est le « tétraèdre » défini par ces points (ce qui nous permet de deviner intuitivement que le 1-simplexe de sommets x_0 et x_1 n'est autre que le segment d'extrémités x_0 et x_1).

A chaque valeur du $(p + 1)$-uplet $(\lambda_0, \lambda_1, ..., \lambda_p)$ de nombres positifs tels que $\sum_{k=0}^{k=p} \lambda_k = 1$, correspond un point du simplexe. La correspondance ainsi définie est bijective, en raison de l'indépendance linéaire des x_j. C'est pourquoi l'on peut dire que les scalaires $\lambda_0, \lambda_1, ..., \lambda_p$ s'appellent les *coordonnées barycentriques* du point $(y^1, y^2, ..., y^n)$ correspondant dans le simplexe.

Pour caractériser, dans le 3-simplexe euclidien, ses faces — au sens usuel — on peut voir que chacune d'entre elles n'est autre que le 2-simplexe défini par les 3 sommets de cette face. Ceci amène à définir, d'une manière plus générale, une *face m-dimensionnelle*, qui est le m-simplexe de m sommets choisis parmi les p du simplexe de départ. On utilise le terme de face — sans préciser plus — lorsque $m = p - 1$; les faces d'un 1-simplexe de sommets x_0 et x_1 sont donc les deux sommets, par exemple. Il y a donc $(p + 1)$ faces pour un p-simplexe euclidien. Ainsi, si dans la définition du p-simplexe, on prend successivement :

$$\lambda_0 = 0, \quad \lambda_1 = 0 ..., \lambda_p = 0,$$

alors les points obtenus décrivent chacune des $(p + 1)$ faces du p-simplexe.

On citera deux propriétés essentielles des simplexes euclidiens.

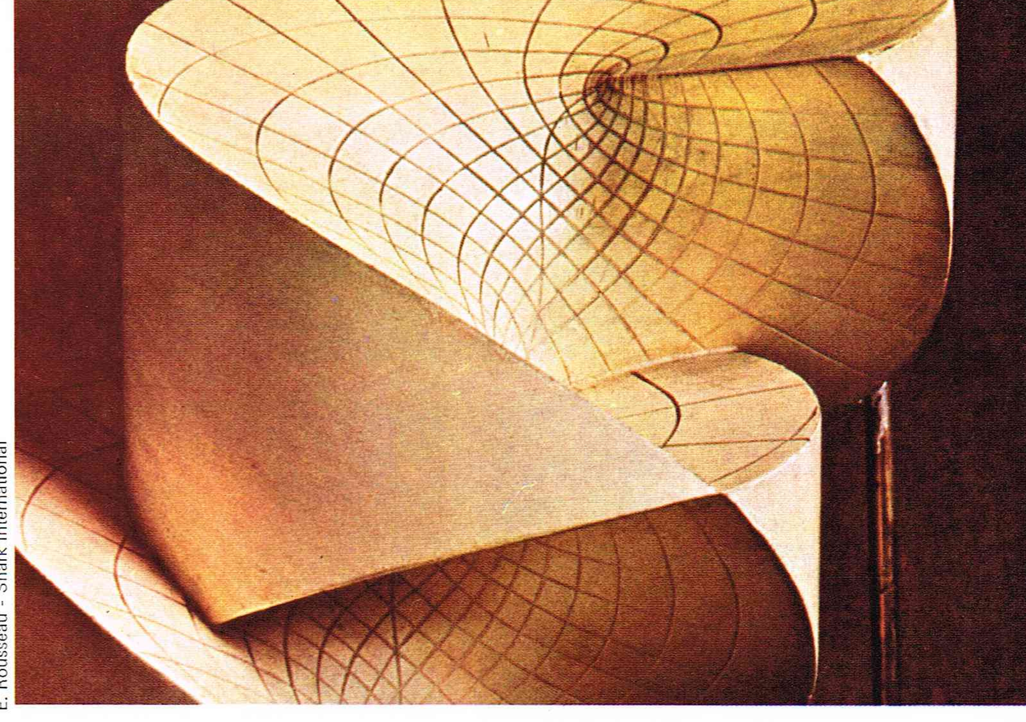
Propriété 1 : tous les points d'un segment joignant deux points d'un simplexe sont dans le simplexe ; autrement dit, tout simplexe euclidien est convexe.

Propriété 2 : le p-simplexe S de sommets $x_0, x_1, ..., x_p$ est exactement formé de l'ensemble des points des segments qui joignent chaque sommet x_k à la face opposée (c'est-à-dire la face obtenue en prenant $\lambda_k = 0$).

Parmi les simplexes *simples*, on peut citer surtout le *p-simplexe euclidien standard* Δ_p (de l'espace à $p + 1$ dimensions), où les sommets sont formés des suites de nombres (0, 0, ..., 0, 1, 0, ..., 0) où seule une coordonnée est non nulle et égale à 1. Dans cet exemple (que l'on peut rapprocher de la base « canonique » de $\mathbb{R}^p$), la recherche des coordonnées barycentriques d'un point du simplexe est très aisée ; on a en fait : $\lambda_k = y^{k+1}$, où y^{k+1} est la $(k + 1)$-ième composante du point y de Δ_p.

Supposons maintenant donné S, un p-simplexe euclidien $(x_0, x_1, ..., x_p)$. Il existe une application linéaire et une seule de Δ_p sur S telle que l'image du i-ième sommet (0, 0, ..., 1, ..., 0), où le terme non nul est à la i-ième place de Δ_p, soit x_{i-1}. Cette application est de plus continue ; on la nomme *p-simplexe singulier associé à S*, et on la note $(x_0, x_1, ..., x_p)$. Plus généralement, on appelle p-simplexe singulier une application continue de Δ_p dans E (espace euclidien de dimension $p + 1$). C'est à l'aide de ces applications que l'on construit maintenant ce qui correspond à notre idée intuitive de courbe limitant une surface (ou encore de surface limitant un solide), c'est-à-dire le bord d'un simplexe.

Dans un p-simplexe, on a pu définir $(p + 1)$ faces, qui sont chacune un $(p - 1)$ simplexe. On pourrait de même penser à « remonter » un simplexe par ses faces ; pour cela, et compte tenu de la nécessité de disposer d'une structure (loi interne, par exemple additive) algébrique, on définit une *p-chaîne* ou chaîne p-dimensionnelle de E comme une combinaison linéaire à coefficients entiers de p-simplexes singuliers de E, soit une expression de la forme $\sum_i n_i s_i$, où seul un nombre fini de n_i sont non nuls. On choisit $n_i \in \mathbb{Z}$, de telle façon que les p-simplexes singuliers de E engendrent ainsi un groupe abélien additif C_p (E), dit *groupe* des p-chaînes de E.



E. Rousseau - Snark International

▲ ***Une surface de Riemann :** les travaux du mathématicien allemand sur les relations entre la théorie des fonctions et la théorie des surfaces sont considérés comme les premières bases de la topologie.*

Le bord d'un simplexe est alors défini par une opération qui, composée avec elle-même, est identiquement nulle. Ceci est naturel ; en effet, le bord d'un triangle (son périmètre), le bord d'un segment (ses extrémités) ont leurs bords respectifs réduits à zéro. L'extrapolation au p-simplexe est donc pleinement justifiée. D'autre part, pour la même raison qui nous a fait définir une p-chaîne de E à partir de p-simplexes singuliers, c'est sur des p-simplexes singuliers que l'on définira l'opération passage au bord. Soit donc $(x_0, x_1, ..., x_p)$ un p-simplexe singulier, on désigne par f_{xk} le $(p - 1)$-simplexe singulier associé à la face opposée au sommet x_k. Alors l'opérateur *bord* d est défini par :

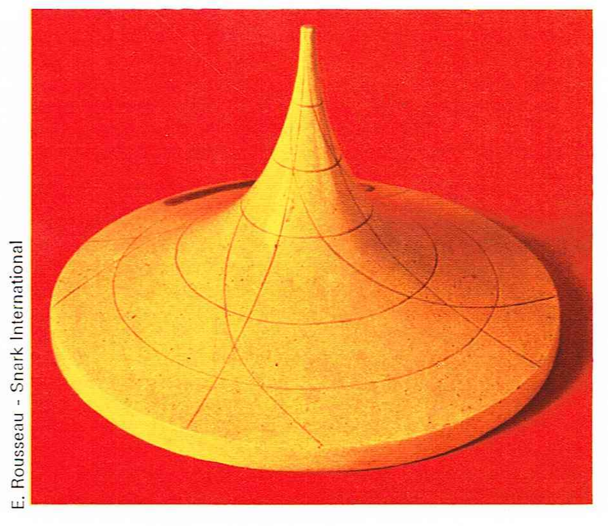
$$d\,(x_0, x_1, ..., x_p) = \sum_{k=0}^{k=p} (-1)^k f_{xk}$$

On en déduit alors que l'on a bien $d^2\,(x_0, x_1, ..., x_p) = 0$.

D'autre part, on définit le bord d'une p-chaîne $\sum_i n_i s_i$ par $\sum_i n_i d s_i$ et, en particulier si C est une 0-chaîne de E, on pose $dC = 0$.

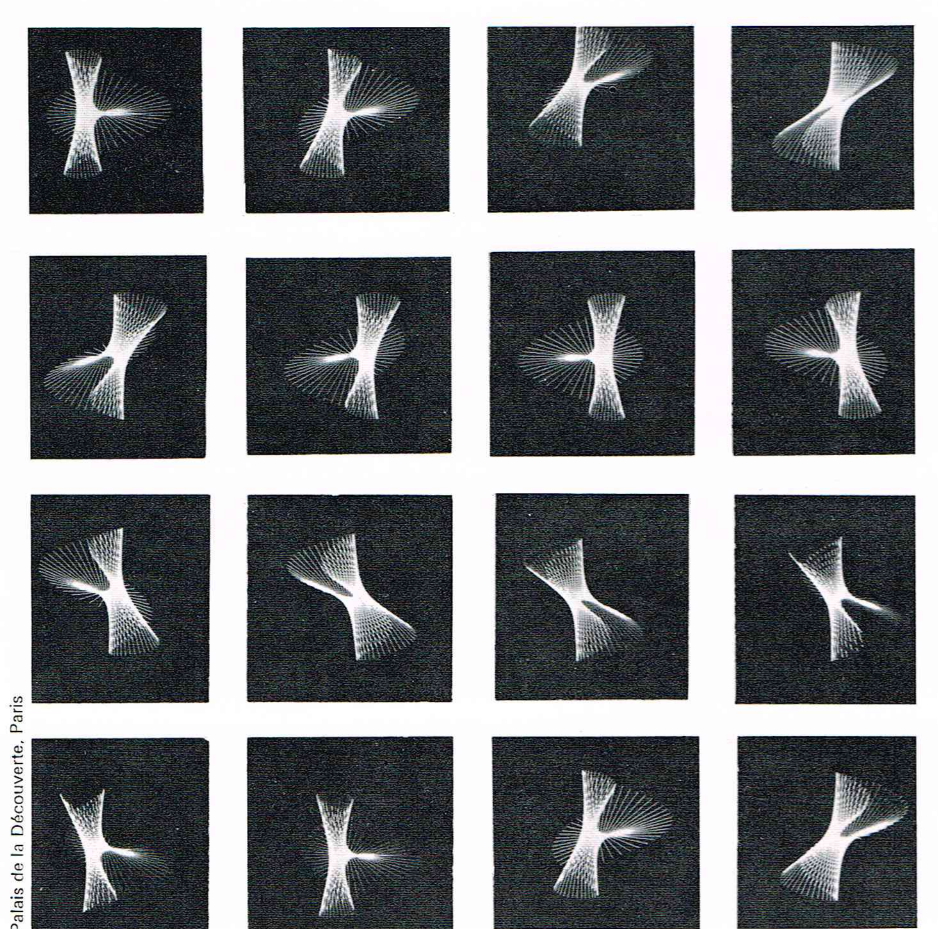
Ceci amène à envisager toutes les p-chaînes dont le bord est nul, afin de retrouver parmi elles celles qui sont des bords. On appelle p-cycle dans un espace E une p-chaîne C telle que $dC = 0$.

On complète cet ensemble de définition en appelant *p-bord* de E toute p-chaîne B de E telle qu'il existe une $(p + 1)$-chaîne C de E vérifiant $B = dC$.



E. Rousseau - Snark International

◀ ***Une surface courbe.***



Palais de la Découverte, Paris

▲ *Rotation d'une* **surface réglée de Möbius.**

Notions d'homologie

Les p-cycles de E forment un sous-groupe de C_p (E), que l'on notera Z_p (E), les p-bords de E forment un sous-groupe de C_p (E), que l'on notera B_p (E). L'opérateur bord, soit d, définit un homomorphisme de C_p (E) dans C_{p-1} (E) pour lequel Z_p (E) et B_p (E) sont respectivement le noyau et l'image. Puisque la relation $d^2 = 0$ est vérifiée pour toute chaîne, il s'ensuit que

$$B_p(E) \subset Z_p(E) ;$$

plus exactement, B_p (E) est un sous-groupe de Z_p (E).

L'idée générale de l'homologie est alors d'effectuer une classification parmi les cycles de E. Non pas pour énumérer ceux d'entre eux qui sont des bords, mais en vue de leur distinction ultérieure.

Soit deux p-cycles A et B de E. S'il existe une $(p+1)$-chaîne C de E telle que $A - B = dC$, on dit alors que A et B sont *homologues* dans Z_p (E), ce qu'on écrit $A \sim B$, car il s'agit bien d'une relation d'équivalence entre p-cycles. Les classes d'équivalences pour cette relation dans Z_p (E) sont appelées *classes d'homologie de dimension p sur* E. C'est évidemment de ces classes d'homologie que l'on peut arriver à faire une liste simple. Mais la relation d'homologie n'est autre que : $A \sim B \Leftrightarrow A - B \in B_p$ (E). Les classes d'homologie qui nous intéressent ne sont donc autres que les éléments du groupe quotient $Z_p(E)/B_p(E)$ que l'on appelle p-*ième groupe d'homologie* de E, ou encore *groupe d'homologie* p-*dimensionnel*, et qu'on désigne par H_p (E).

L'étude de la structure de ces groupes, de leurs générateurs et des liens entre ceux-ci permettra de déterminer les bords dans E.

On dit que les p-cycles $C_1, C_2, ..., C_m$ sont linéairement indépendants dans E si la seule combinaison linéaire $a_1C_1 + a_2C_2 + ... a_mC_m$ à coefficients entiers, homologue à zéro [donc bord d'un $(p+1)$ cycle], est celle pour laquelle $a_1 = a_2 = ... = a_m = 0$. C'est le plus grand nombre de p-cycles linéairement indépendants d'une variété V que l'on appelle le *nombre de connexions*, ou *nombre de Betti*, de dimension p de V.

Un cas particulier est celui d'un p-cycle C non homologue à zéro, mais pour lequel il existe un nombre $k \in \mathbb{N}$ tel que $kC \sim 0$; alors le plus petit entier k qui réalise cette condition est appelé le *coefficient de torsion* de dimension p de V.

Le p-ième groupe d'homologie d'une variété V est somme finie de groupes cycliques, et le nombre de connexions de dimension p n'est autre que le nombre des groupes cycliques infinis intervenant dans H_p (V) ; tandis que les ordres des groupes finis sont des multiples du coefficient de torsion de dimension p de V. Les *relations de dualité de Poincaré* montrent l'égalité, pour une variété compacte et orientable de dimension p, des nombres de connexions de dimensions n et $p - n$, ce pour toutes les valeurs de n entières comprises entre 0 et p ; ainsi que l'égalité des coefficients de torsion de dimensions n et $p - n - 1$ pour les mêmes valeurs de n.

Les nombres de connexions possèdent la propriété fondamentale d'être invariants pour tout homéomorphisme de V.

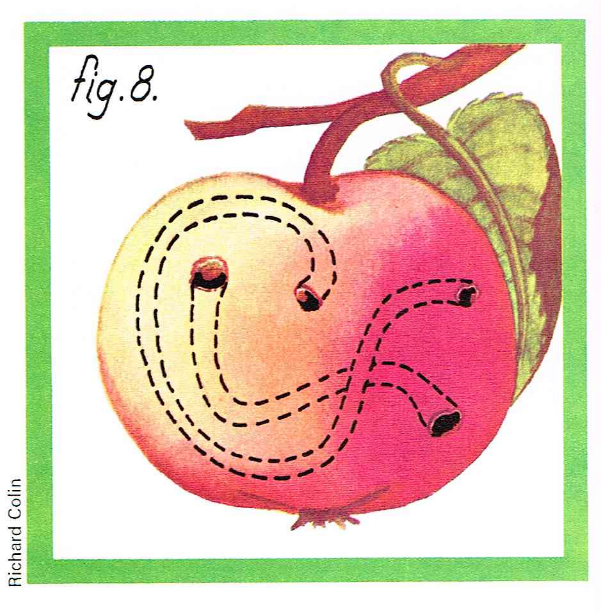
Le *théorème de dualité d'Alexander* pour les sommes finies de simplexes globalise un certain nombre de résultats tels que le théorème de Jordan sur la découpe en 2 parties du plan euclidien par une ligne simple fermée, ou le *théorème de Brouwer* (qui généralise le résultat précédent à une variété compacte de dimension $p - 1$ dans un espace de dimension p), ou encore le théorème de l'invariance de la dimension topologique.

Si K désigne un nombre fini ou une infinité dénombrable de simplexes tels que deux simplexes quelconques aient soit un côté commun, soit une intersection vide, et que tout point de K appartienne à un nombre fini de simplexes (alors on dit que K est un complexe, de dimension n si tous les simplexes ont une dimension au plus égale à n), alors le nombre de connexions de dimension r de k est égal au nombre de connexions de dimension $n - r - 1$ de l'ensemble $\complement K$.

La généralisation des notions précédentes permet de décrire des cas tels que celui d'une pomme rongée par deux vers, où les canaux tracés ne se rencontrent pas et ne forment donc jamais une *coupure* du fruit, mais néanmoins délimitent une *bordure* des chemins suivis (*fig.* 8). Soit donc E un espace et F un sous-espace. Un *p-cycle relatif de E modulo F* est une p-chaîne C de E telle que dC soit une $(p-1)$ chaîne de F. De manière analogue, un p-bord relatif de E modulo F est une p-chaîne B de E, telle qu'il existe une p-chaîne s de F et une $(p+1)$ chaîne t de E vérifiant $dt = s + B$.

Les p-cycles relatifs de E modulo F forment un sous-groupe de C_p (E) qu'on note Z_p (E, F) ; les p-bords relatifs de E modulo F forment un sous-groupe de C_p (E), noté B_p (E, F), et on voit que B_p (E, F) est un sous-groupe de Z_p (E, F), ce qui permet de définir le p-ième groupe d'homologie relatif de E, modulo F, soit

$$H_p(E, F) = Z_p(E, F) / B_p(E, F).$$



Richard Colin

▶ *Figure 8 :* ***généralisation*** *des* **notions d'homologie** ***(voir développement dans le texte ci-contre).***

Ce groupe contient comme éléments les classes de p-cycles relatifs de E modulo F, pour la relation d'équivalence définie par : deux éléments de Z_p (E, F) sont équivalents si leur différence est un élément de B_p (E, F).

Le calcul des groupes d'homologie et d'homologie relative permet ainsi une connaissance plus précise des propriétés topologiques de toutes les surfaces n-dimensionnelles. Un outil très précieux dans les calculs de certains groupes d'homologie est le *théorème d'excision* qui montre que l'homomorphisme :

$$H_p (E - U, F - U) \rightarrow H_p (E, F)$$

induit par l'application inclusion est un isomorphisme dès que U est un sous-ensemble de F tel que $\bar{U} \subset \mathring{F}$ — avec F sous-espace de E — et ce pour toutes les valeurs de p. Autrement dit, il est possible de supprimer *(d'exciser)* certains sous-ensembles dans F, sans pour autant que l'homologie relative (et là réside un grand intérêt de cette généralisation) de E modulo F soit transformée.

Notions d'homotopie

Un aspect très intéressant de la topologie algébrique est la théorie de l'homotopie des arcs.

Soit I l'intervalle formé [0, 1] de $\mathbb{R}$, muni de la topologie induite par celle de $\mathbb{R}$. Un *arc* α d'un espace topologique E est une application continue de I dans E, c'est-à-dire une fonction continue α (t) $(0 \leqslant t \leqslant 1)$ dont les valeurs sont des points de E. Ordinairement, pour tout arc α on a une courbe image, lieu des points de E qui sont images par α d'un point de I ; mais l'arc ne se réduit pas à un tel lieu, il est plutôt une de ses représentations paramétriques. Ainsi, dans $\mathbb{R}^2$, les arcs : $x_k = kt^2$ et $x_k = kt$ ont la même courbe image — le segment d'extrémités (0, 0) et (1, 2) — mais sont différents. En général, si f est un homéomorphisme de I dans lui-même, l'application composée de f et α, soit $\alpha \circ f$, a la même courbe image que l'arc α. Puisque l'on n'a pas demandé que α soit injective, un point de E peut être image de plusieurs points de I. Dans cet ordre d'idée, Peano a construit un arc de $\mathbb{R}^2$ dont la courbe image est un carré. Les points α (0) et α (1) sont appelés les *extrémités* de l'arc α ; lorsque $\alpha (0) = \alpha (1)$, on dit que l'on a un *lacet*, et si α (t) est constant, on parle d'arc nul.

Dans l'ensemble des arcs de E, on introduit une opération non définie partout : le *produit homotopique*. Étant donné deux arcs de E, α et β, leur produit $\alpha\beta$ (dans cet ordre) est défini si $\alpha (1) = \beta (0)$ et est donné par :

$$\alpha\beta\ (t) = \begin{cases} \alpha\ (2\,t) \text{ pour } 0 \leqslant t \leqslant \dfrac{1}{2} \\ \beta\ (2\,t - 1) \text{ pour } \dfrac{1}{2} < t \leqslant 1 \end{cases}$$

Une telle application est continue, et manifestement on a :

$$\alpha\beta\ (0) = \alpha\ (0) \qquad \text{et} \qquad \alpha\beta\ (1) = \beta\ (1)$$

Intuitivement, on peut interpréter t comme une variable intermédiaire ; alors α et β sont les lois horaires de deux mouvements dans E, qui se développent dans l'unité de temps I ; $\alpha\beta$ est alors une nouvelle loi horaire qui se réalise en effectuant dans la première moitié de I le mouvement α à vitesse double, et dans la seconde moitié, le mouvement β également à double vitesse. Le produit homotopique définit dans l'ensemble des arcs de E une structure algébrique.

Convenons maintenant de définir dans un tel ensemble une relation d'équivalence, l'*homotopie*. On dira que deux arcs α et γ de E sont homotopes (*fig.* 9) s'il existe une application continue H de I^2 dans E telle que :

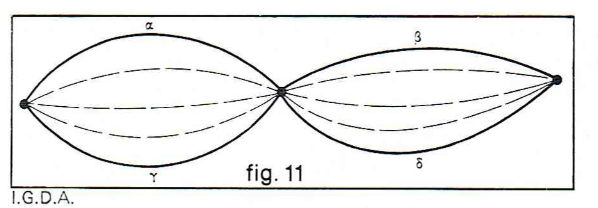
$$\mathbf{H_1)} \qquad \begin{cases} H\ (t, 0) = \alpha\ (t) \\ H\ (t, 1) = \gamma\ (t) \end{cases}$$

Il est ensuite utile de considérer une relation plus forte, dite d'*homotopie relative* ($\sim$), qu'on a lorsque α, γ, H satisfont en plus de $\mathbf{H_1}$ à :

$$\mathbf{H_2)} \qquad \begin{cases} \alpha\ (0) = \gamma\ (0) = H\ (0, u) \\ \alpha\ (1) = \gamma\ (1) = H\ (1, u) \end{cases} \qquad (\textit{fig.}\ 10)$$

D'un point de vue intuitif, on peut dire que α et β sont homotopes lorsque l'on peut transformer par continuité l'un en l'autre ; en posant $u = u_0$, H (t, u_0) est alors un des arcs qui réalise un tel passage.

Inversement, si l'on pose $t = t_0$, H (t_0, u) est l'arc *décrit* par un point dans la transformation. Dans l'homotopie relative, on désire que tous les arcs H (t, u_0) — et de là en particulier α et γ — aient les mêmes extrémités.



I.G.D.A.

◄ *Figure 11 :* ***le produit homotopique est compatible avec l'homotopie relative.***

Il est facile de vérifier que les deux relations sont des relations d'équivalence. En outre, le produit homotopique est compatible avec l'homotopie relative, c'est-à-dire que si $\alpha \sim \gamma$ et $\beta \sim \delta$ et si $\alpha\beta$ existe, alors $\gamma\delta$ est aussi défini et $\alpha\beta \sim \gamma\delta$ (*fig.* 11).

On appelle *classes d'homotopie* d'un point x de E les classes d'équivalence de l'homotopie relative dans l'ensemble des lacets ayant leurs extrémités en x. Désignons par $\mathcal{C}_\alpha$ la classe à laquelle appartient le lacet α ; on peut alors définir une opération interne dans l'ensemble $\mathcal{C}$ des classes $\mathcal{C}_\alpha$ comme suit :

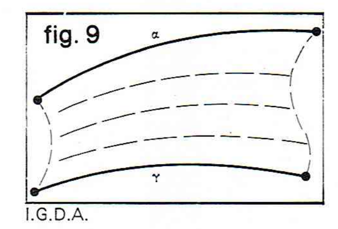
$$\mathcal{C}_\alpha\, \mathcal{C}_\beta = \mathcal{C}_{\alpha\beta}$$

(en fait, si $\gamma \in \mathcal{C}_\alpha$ et $\delta \in \mathcal{C}_\beta$, alors $\mathcal{C}_{\gamma\delta} = \mathcal{C}_{\alpha\beta}$, et par conséquent il en résulte des opérations qui dépendent seulement des classes $\mathcal{C}_\alpha$, $\mathcal{C}_\beta$ et non des arcs α, β). Cette opération est toujours définie et fait de $\mathcal{C}$ un groupe que l'on appelle *groupe d'homotopie de x*.

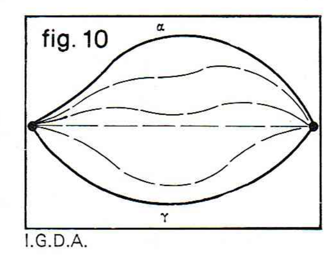
On dira que E est *connexe par arcs* si, pour tout couple (x, y) de points de E, il existe toujours un arc de E d'extrémités x et y. Si E est connexe par arcs, les groupes d'homotopie de ses points sont isomorphes à un même groupe que l'on appelle *groupe fondamental*, ou premier groupe d'homotopie, ou *groupe de Poincaré* de E, et que l'on note π (E). Deux espaces homéomorphes ont le même groupe fondamental.

Le plan $\mathbb{R}^2$, l'espace $\mathbb{R}^3$, une surface sphérique, bien que n'étant pas homéomorphes, ont le même groupe d'homotopie : en fait, deux arcs avec les mêmes extrémités sont relativement homotopes, et de là, à tout point on peut associer une seule classe d'homotopie ; alors π (E) se réduit à un groupe à un élément : l'élément neutre.

Dans la couronne circulaire, on a inversement des lacets non homotopes au lacet nul (*fig.* 12-1), et le groupe d'homotopie est isomorphe au groupe additif des entiers relatifs ($\mathbb{Z}^+$). Sur le tore (*fig.* 12-2) on a deux types de lacets non homotopes au lacet nul et non réductibles l'un à l'autre ; c'est pourquoi, dans ce cas, π (E) est isomorphe au groupe additif des nombres complexes $x + iy$ tels que x et y soient entiers.

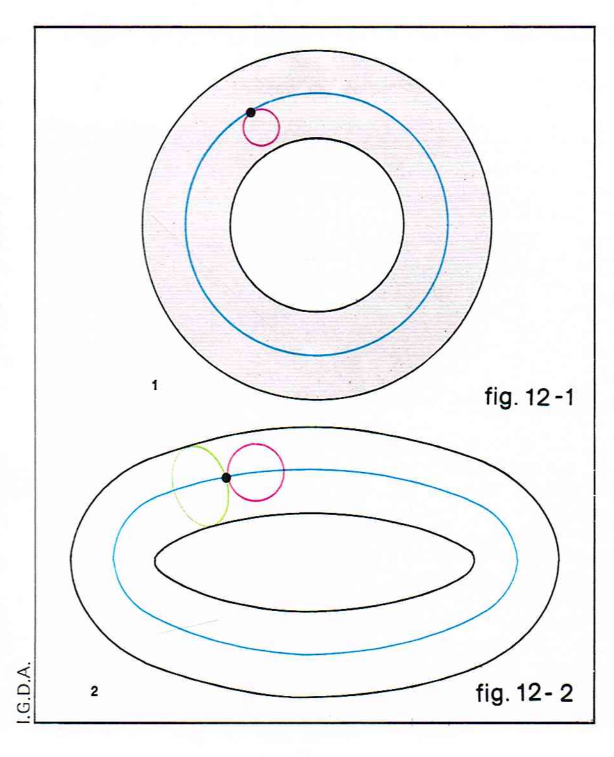


I.G.D.A.



I.G.D.A.

▲ *En haut, figure 9 :* ***l'homotopie de deux arcs.*** *Ci-dessus, figure 10 :* ***l'homotopie relative.***



I.G.D.A.

◄ ***Figure 12 : lacets homotopes au lacet nul (en rouge) et non homotopes au lacet nul (en bleu et en vert) illustrés par l'exemple d'une couronne circulaire (1) et par celui d'un tore (2).***

La notion d'homotopie s'étend aux applications continues d'un espace donné X dans un espace donné E : l'homotopie entre les applications f et g est réalisée par une application continue H de X × I dans E avec

$$\begin{cases} H(x, 0) = f(x) \\ H(x, 1) = g(x) \qquad \text{pour tout } x \in X \end{cases}$$

En particulier, appelons S_n la sphère n-dimensionnelle, lieu des points de $\mathbb{R}^{n+1}$ tels que $\sum_{i=1}^{i=n+1} x_i^2 = 1$; on peut étendre les considérations précédentes et obtenir un nouveau groupe π_n (E).

La théorie des applications homotopes étend considérablement la notion de continuité. Certains résultats combinent homotopie et groupes d'homologie ; le principal d'entre eux est :

soit X et E deux espaces topologiques, Y et F respectivement deux sous-espaces. Soit f et g deux applications de X dans E, telles que Y soit appliqué dans F, et homotopes. Si l'on désigne par f_* et g_* les morphismes de H_p (X, Y) dans H_p (E, F) induits par f et g, alors $f_* = g_*$, et ce pour toute valeur de p.

En conséquence, si $f : X \to E$ et $g : E \to X$ continues sont telles que $f(Y) \subset F$ et $g(F) \subset Y$ et si $f \circ g$ est homotope à l'identité de E et $g \circ f$ homotope à l'identité de X, alors le morphisme f_* induit par f,

$$f_* : H_p(X, Y) \to H_p(E, F)$$

est un isomorphisme pour tout $p \in \mathbb{N}$.

Il résulte encore de là que si E et F sont des espaces topologiques, $f : E \to F$ et $g : F \to E$ des applications continues, alors, si $f \circ g$ et $g \circ f$ sont homotopes aux applications identités de F et de E respectivement, alors :

$$f_* : H_p(E) \to H_p(F)$$

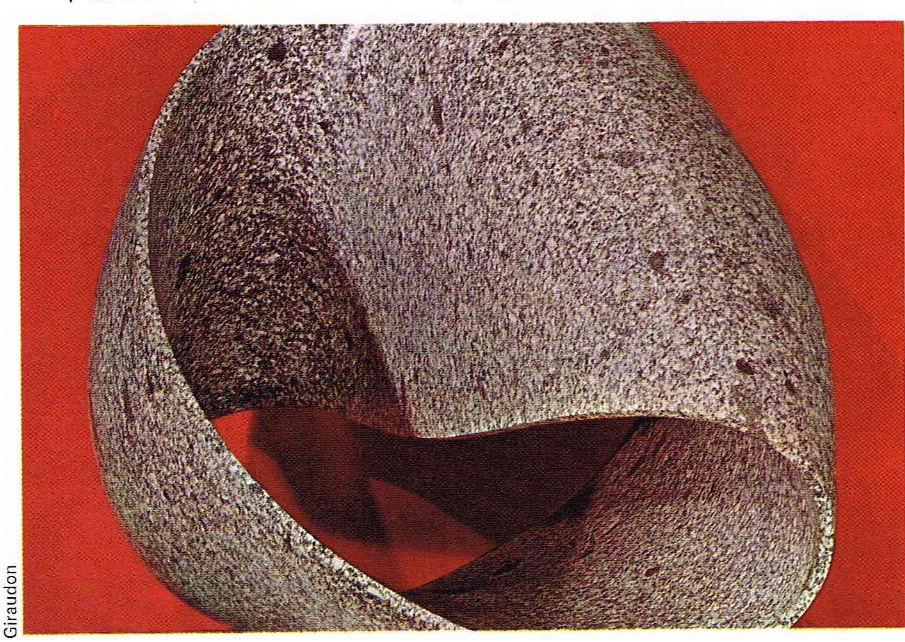
induit par f est un isomorphisme, pour toute valeur de p.

L'homotopie et l'homologie permettent ainsi de distinguer, du point de vue topologique, des espaces que l'on pourrait croire homéomorphes alors qu'ils ne le sont pas.

BIBLIOGRAPHIE

BOURBAKI N., *Topologie générale*, ch. I à IV, Hermann. - GODBILLON C., *Éléments de topologie algébrique*, Hermann. - GODEMENT R., *Topologie algébrique et Théorie des faisceaux*, Hermann. - LELONG-FERRAND J., *Géométrie différentielle*, Masson. - LICHNEROWICZ A., *Théorie globale des connexions et des groupes d'holonomie*, Cremonese, Rome. - SPANIER E. H., *Algebric Topology*, Mac Graw Hill. - WALLACE A. H., *Introduction à la topologie algébrique*, Gauthier-Villars. - ZISMAN M., *Topologie algébrique élémentaire*, Armand Colin.

▼ ***Le ruban de Möbius.***



Giraudon

STATISTIQUE ET PROBABILITÉS

Généralités

Chacun sait ce que signifie le concept de statistique. On fait la statistique des accidents de la route, du nombre de naissances ou des gagnants à la Loterie nationale. Cela veut dire que l'on compte le nombre des accidents de la route qui se sont produits dans un laps de temps donné, ou celui des naissances survenues dans la population de tel ou tel pays, ou enfin celui des personnes qui ont eu la chance de gagner telle ou telle somme à la Loterie.

La statistique est donc simplement une technique, une sorte de comptabilité des événements qui se sont produits dans le passé.

La notion de probabilité est liée à cette technique, mais elle est tournée vers le futur. Si je vous demande : que va-t-il vous arriver cette année ? La réponse à ma question contiendra la notion de probabilité. A la base de cette notion on trouve une supposition fondamentale *a priori :* à savoir que, toutes choses égales par ailleurs, les divers événements dont on a fait la statistique sont équivalents, en ce sens que chacun d'eux a une égale probabilité de se reproduire dans le futur comme il s'est produit dans le passé. Ainsi, si une partie de pile ou face a donné jusqu'à maintenant 60 « piles » et 40 « faces », on considérera que chacun de ces 100 événements a la même probabilité de se reproduire au 101ᵉ essai, et que par conséquent, pour ce nouvel essai, on a 60 % de chances d'obtenir « pile » ou simplement que la probabilité d'obtenir « pile » est 0,6. Cette équivalence des divers événements passés n'est, soulignons-le, rien de plus qu'une hypothèse de travail, que rien ne justifie, si ce n'est son efficacité pratique.

La probabilité intrinsèque

La notion empirique de probabilité qui résulte d'une statistique est, naturellement, toujours sujette à révision. Chaque fois qu'une nouvelle expérience est faite, le résultat de cette expérience modifie, en général, les proportions entre les divers résultats survenus, et donc la « probabilité » de chacun d'eux pour un essai ultérieur. Le concept de probabilité s'affranchit de cette contrainte et acquiert une indépendance par rapport à l'expérience, si nous supposons qu'il existe une grandeur appelée *probabilité intrinsèque,* liée à chaque événement susceptible de se produire, grandeur dont les statistiques donnent une évaluation plus ou moins approchée.

Ainsi, nous pouvons attribuer, pour des raisons de symétrie évidentes, à l'événement « pile » la probabilité intrinsèque 0,5, et affirmer que l'évaluation de cette probabilité au terme de 100 parties (0,6 dans notre exemple) ne constitue qu'une évaluation grossière de la probabilité intrinsèque.

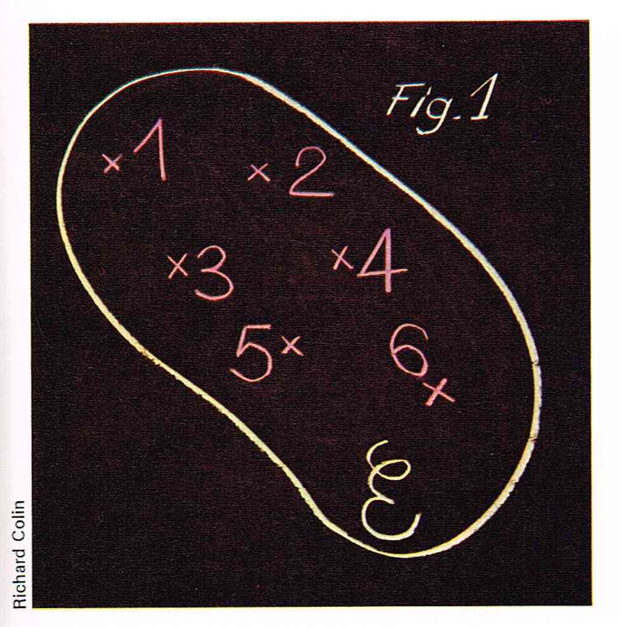
Il est possible de démontrer, sous des conditions très générales, que l'évaluation des probabilités intrinsèques (auxquelles nous réserverons par la suite le terme de probabilité) au moyen d'une statistique est d'autant meilleure que le nombre d'événements enregistrés dans la statistique est plus grand.

L'espace d'événements possibles

On se demande souvent quelle est la probabilité de tel événement, alors même qu'aucune expérience n'a été faite pour donner une statistique à cet égard. On considère alors l'ensemble des événements possibles, ou plus précisément, l'ensemble des *résultats* physiquement distincts pouvant résulter d'une *épreuve* ou d'un *essai.* Cet ensemble forme l'*espace des événements possibles.* Il est parfois utile de se représenter cet espace. La *figure* 1 illustre, par exemple, l'espace $\mathcal{E}$ des événements possibles pour un tirage de dés.

L'hypothèse d'égale probabilité a priori

Afin de répondre à une question « probabilistique » du type : quelle est la probabilité de tirer un chiffre pair en jetant le dé ? on énumère les événements favorables, c'est-à-dire, dans ce cas, les événements « sortie d'un 2 », « sortie d'un 4 », « sortie d'un 6 ». Le rapport entre le nombre des événements favorables (N_f) et le nombre total





▲ *A gauche, figure 1 : l'espace* ℰ *des événements possibles pour un tirage de dés. A droite, l'évaluation de la probabilité est fondée sur une supposition* a priori *: celle de l'équivalence de chacun des événements élémentaires qui forment l'espace des événements possibles.*

des événements possibles (N) est, par hypothèse, la probabilité en question. On écrit :

(1) $$P = \frac{N_f}{N}.$$

Ici encore, l'évaluation de la probabilité est fondée sur une supposition *a priori :* celle de l'équivalence de chacun des événements élémentaires qui forment l'espace des événements possibles. C'est cette hypothèse, par exemple, qui nous a permis d'attribuer la probabilité 0,5 à chacun des événements « pile » et « face », parce qu'il n'y a dans ce cas que deux événements possibles, et qu'ils sont considérés comme équivalents.

La fécondité de cette hypothèse, qu'on appelle souvent l'*hypothèse d'égale probabilité a priori*, est l'unique justification qui puisse en être donnée. Dans certains cas, la fécondité de cette hypothèse résulte de la propre volonté des observateurs. Dans les jeux, par exemple, on cherche délibérément à créer les conditions de son application, sans jamais y parvenir parfaitement ; car, comme on peut le remarquer, chaque face d'un dé diffère des autres par un détail quelconque, sans quoi il ne serait pas possible de les distinguer. En fait, on rencontre dans la nature d'innombrables cas où l'*hypothèse d'égale probabilité a priori* se révèle d'une fécondité surprenante, à tel point que l'on peut se demander si l'on a bien compris toute la portée ou toute la signification d'un tel succès.

Probabilités composées : lois de composition

Événements élémentaires et événements composés

Considérons un espace d'événements élémentaires, par exemple celui des résultats possibles du tirage d'un dé. A chaque question probabilistique on peut associer l'ensemble des événements élémentaires favorables, qui, conformément à la relation **(1)**, permet de calculer la probabilité en question.

Un tel ensemble d'événements élémentaires favorables est aussi appelé, par extension, *un événement.* Ainsi, nous demanderons : Soit A le sous-espace de ℰ tel que..., quelle est la probabilité pour que l'événement A se produise ? Dans le jeu de dés, on peut définir, par exemple, l'événement A comme « sortie d'un chiffre impair », l'événement B « sortie du chiffre un », l'événement C « sortie des chiffres 5 ou 6 », D « sortie d'un chiffre pair », etc. Graphiquement (*fig.* 2), à chacun de ces événements correspond une région de ℰ contenant exclusivement les événements élémentaires favorables.

Quelle que soit la définition de l'événement E, la relation **(1)** indique que

(2) $$0 \leqslant Pr\{E\} \leqslant 1.$$

On peut construire, à partir des événements ainsi définis, d'innombrables autres événements ; « A et B », par exemple, est le sous-espace contenant l'ensemble des événements élémentaires favorables à la fois à A et à B ; « A ou B » est l'événement formé par les événements favorables, soit à A, soit à B. Appelant E et F, respectivement, les événements précédents (*fig.* 3), on écrit :

$$E = A \cap B$$
$$F = A \cup B.$$

— Si A et B s'excluent mutuellement, c'est-à-dire s'ils n'ont aucun événement élémentaire en commun, E est vide :

$$Pr\{E\} = 0.$$

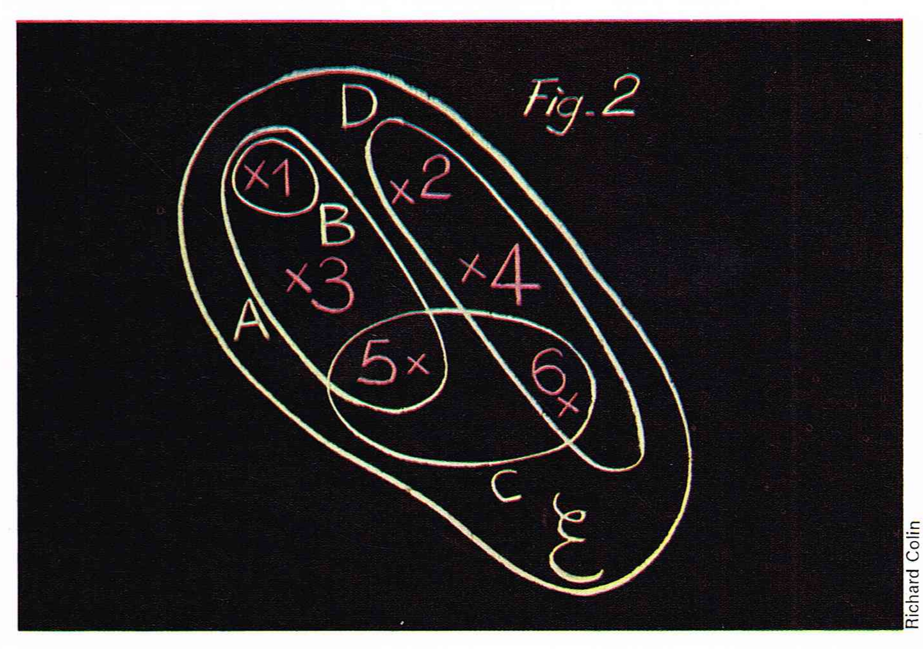
— Si $A \cup B = \mathcal{E}$, et si A et B s'excluent mutuellement, on dit que A et B sont complémentaires. On écrit $B = \bar{A}$. Il est clair que $Pr\{A \cup \bar{A}\} = Pr\{\mathcal{E}\} = 1$.

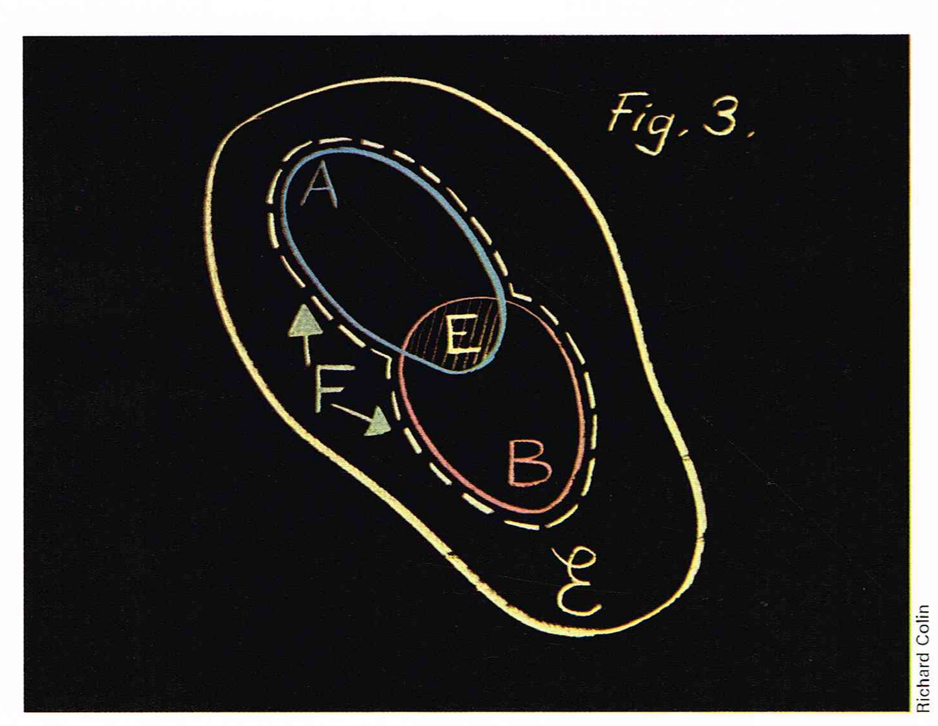
La signification de ces termes devient évidente si l'on considère des exemples particuliers, comme celui du jeu de dés cité plus haut. Dans la *figure* 2 les événements B et C sont exclusifs, les événements A et D sont complémentaires. De façon générale, les événements élémentaires de n'importe quel espace des événements possibles sont toujours mutuellement exclusifs.

Probabilités conditionnelles

Il est parfois important de considérer un événement particulier comme définissant un nouvel espace des événements possibles. C'est le cas, par exemple, pour la question

▼ *Figure 2 : à chacun des événements A, B, C et D, correspond une région de* ℰ *contenant exclusivement les événements élémentaires favorables.*





Richard Colin

▲ *Figure 3 : voir développement dans le texte.*

suivante : « quelle est la probabilité d'avoir tiré un 1 si, pour une raison quelconque, nous savons que le résultat du tir est impair ? » Une telle probabilité est appelée *probabilité conditionnelle*, et l'événement G correspondant se met sous la forme :

$$G = B/A.$$

Pr (B/A) est donc la probabilité pour que B se produise, sachant que A s'est effectivement produit. On a :

$$Pr\,\{G\} = Pr\,\{B/A\} = \frac{N_f}{N_A}$$

Lois de composition

La relation précédente donne immédiatement le moyen d'établir deux théorèmes fondamentaux de la théorie des probabilités. Étant donné que :

$$Pr\,\{B/A\} = \frac{N_f}{N_A}$$

peut s'écrire :

$$Pr\,\{B/A\} = \frac{\frac{N_f}{N}}{\frac{N_A}{N}} = \frac{Pr\,\{A \cap B\}}{Pr\,\{A\}}$$

on obtient :

(3) $$Pr\,\{A \cap B\} = Pr\,\{A\} \times Pr\,\{B/A\}.$$

D'autre part, le comptage des événements favorables pour l'union de deux événements donne immédiatement :

$$Pr\,\{A \cup B\} = Pr\,\{A\} + Pr\,\{B\} - Pr\,\{A \cap B\}$$

expression dans laquelle le dernier terme compense le fait que les événements de la région commune à A et B sont contenus dans chacun des termes précédents et sont donc comptés deux fois dans la somme de ces termes. Si A et B sont exclusifs, $Pr\,\{A \cap B\}$ est nulle.
La formule précédente se réduit alors à :

(4) $Pr\,\{A \cup B\}$, A et B exclusifs $= Pr\,\{A\} + Pr\,\{B\}$

Tribus et espaces probabilisables

La possibilité de construire, à partir des éléments d'un espace $\mathcal{E}$ d'événements possibles, une multitude d'autres événements (grâce aux lois de composition) conduit les mathématiciens à associer à l'espace $\mathcal{E}$, supposé non vide, les espaces E, dits *tribus de parties de* $\mathcal{E}$, tels que

1) $\mathcal{E} \in E$
2) $e \in E \Rightarrow \bar{e} \in E$
3) Si $\{e_k\}_{k=1,\ldots,n} \in E$, alors $\{e_1 \cup e_2 \ldots \cup e_N\} \in E$.

Le couple $\{\mathcal{E}, E\}$ formé par l'ensemble $\mathcal{E}$ et une tribu de parties de $\mathcal{E}$ est appelé un *espace probabilisable*. On réserve généralement le terme d'*épreuve* pour les événements élémentaires de $\mathcal{E}$, tandis que le terme *événement* s'applique à tout élément $e \in E$. $\{\mathcal{E}\}$ est l'événement certain, $\{\varnothing\}$ (partie vide), en tant qu'élément de E, est l'événement impossible.

Dans ce texte, toute application $Pr : E \to [0, 1]$ telle que $Pr\,\{\mathcal{E}\} = 1$ et telle que pour tout couple exclusif $e_i, e_j \in E$ (tel que $e_i \cap e_j = \varnothing$), $Pr\,\{e_i \cup e_j\} = Pr\,\{e_i\} + Pr\,\{e_j\}$ peut définir la *probabilité Pr sur la tribu E* (en mathématiques, par conséquent, la probabilité peut se définir en dehors de la restriction de l'hypothèse d'égale probabilité *a priori* pour les épreuves — ou événements élémentaires de $\mathcal{E}$).

Variables aléatoires et distributions de probabilité

Le concept de variable aléatoire

Supposons que nous ayons un espace des événements possibles $\mathcal{E}$ et que nous décidions d'attribuer à une variable algébrique X la valeur x_1 si le tirage au sort sur $\mathcal{E}$ réalise l'événement e_1, la valeur x_2 si le tirage au sort réalise l'événement e_2, etc. Par exemple, $\mathcal{E}$ est l'espace des événements possibles au jeu de roulette et X prend la valeur de la perte ou du gain que nous réaliserons si la bille s'arrête sur telle ou telle case.

On dit que X est une *variable aléatoire*. Dans ce cas, elle est attachée à l'espace des événements $\mathcal{E}$, ou, plus précisément, elle est définie par une application de l'espace des événements possibles sur le corps des réels :

$$\{e_n\} \to X = x_n \in R.$$

Notons que cette application (définie *a priori*) peut être, ou non, biunivoque. Par exemple, pour le jeu de dés, on peut définir X comme la variable qui prend les valeurs 1, 2, ..., 6 quand le dé indique respectivement un 1, un 2 ou un 6 ; mais X pourrait aussi être définie comme la variable, prenant la valeur 1 quel que soit le numéro indiqué sur le dé, excepté quand celui-ci est un 6, auquel cas X prend la valeur 2, etc.

Une variable aléatoire X est *discrète* si les valeurs possibles pour $X = x_1 \ldots, x_n$ forment un ensemble de valeurs discrètes, ou *continue* si X peut prendre une valeur quelconque de l'intervalle $[x_a, x_b] \in R$ (c'est le cas, par exemple, de la variable aléatoire définie par les positions possibles d'une aiguille sur le cadran d'un appareil de mesure).

Loi de probabilité se rapportant à une variable aléatoire discrète

Définition

L'union de tous les événements élémentaires auxquels est associée la même valeur x_j de X définit un événement aléatoire a_j. Par définition, la probabilité pour que X prenne la valeur x_j est égale à la probabilité de réalisation de l'événement a_j.

Si l'application de $\mathcal{E}$ sur R est biunivoque, a_j se réduit à un seul événement élémentaire e_j, mais ceci n'est pas toujours le cas. Dans les deux exemples tirés du jeu de dés cités plus haut, la probabilité pour que X prenne la valeur 1, par exemple, est respectivement égale à la probabilité de réalisation des événements :

— le dé indique un 1 (= 1/6) ;
— le dé indique un chiffre quelconque à l'exception du 6 (= 5/6).

La probabilité pour que X prenne la valeur x_j définit une application F (x_j), que nous allons appeler *loi*, ou en suivant un usage ancien quoique impropre, la *distribution de probabilité* de la variable aléatoire X. Cette application est limitée, puisque les probabilités vérifient la relation **(2)**. Nous écrirons, en utilisant une notation évidente :

(5) $$\begin{cases} Pr\,(X = x_j) = F\,(x_j) \\ 0 \leqslant F\,(x_j) \leqslant 1 \\ \Sigma\, F\,(x_j) = 1 \end{cases}$$

où la dernière relation exprime le fait que la variable aléatoire X prend nécessairement l'une quelconque des valeurs possibles x_j.

La distribution hypergéométrique

Comme exemple d'une distribution de probabilité se rapportant à une variable discrète, nous prendrons la distribution hypergéométrique, liée au problème suivant : soit une population U de n objets. Supposons que nous choisissons au hasard r objets parmi cette population U. Le nombre de lots de r objets, sans distinguer leur ordre et sans prendre deux fois le même objet, qui puisse se faire à partir de la population U est $\binom{n}{r} = \frac{n!}{[r!\,(n-r)!]}$.

Considérons maintenant chacun de ces lots comme un événement élémentaire d'un espace $\mathcal{E}$ des événements possibles, et supposons, en outre, que les n objets de la population U sont de deux types : n_1 sont du type A, et les $n - n_1$ restants du type B. Appelons K la variable aléatoire qui prend la valeur définie, à chaque tirage, par le nombre k d'objets du type A présent dans le lot. Quelle est la distribution de probabilité F (k) se rapportant à K?

Pour rendre notre discussion plus concrète, nous pouvons supposer que U est une urne contenant n_1 boules blanches et $n - n_1$ boules noires. Nous tirons au hasard r boules hors de l'urne. F (k) représente dans ce cas la probabilité pour que le lot tiré contienne k boules blanches et $r - k$ boules noires (*fig.* 4).

La probabilité en question est, d'après **(1)**.

$$Pr = \frac{N_f}{N} = \frac{N_f}{\binom{n}{r}}$$

où N_f est le nombre de lots de r boules contenant exactement k boules blanches. On a $\binom{n_1}{k}$ manières de choisir ces k boules blanches parmi les n_1 boules blanches disponibles. Chaque choix peut être combiné avec $\binom{n-n_1}{r-k}$ manières de choisir les $r - k$ boules noires ; de sorte que :

$$\textbf{(6)} \qquad F_{(k)} = \frac{\binom{n_1}{k}\binom{n-n_1}{r-k}}{\binom{n}{r}}$$

Il est facile de démontrer, en se servant des définitions des coefficients binominaux, que F (k) peut aussi s'écrire :

$$F_{(k)} = \frac{\binom{r}{k}\binom{n-r}{n_1-k}}{\binom{n}{n_1}}$$

La *figure* 5 représente la distribution F (k) dans le cas particulier $r = 4$, $n_1 = 5$, $n = 10$. Dans cet exemple, k peut prendre seulement les valeurs : 0, 1, 2, 3 et 4. Les probabilités respectives sont : 5/210, 50/210, 100/210, 50/210, 5/210. Il est clair que les relations **(5)** sont satisfaites dans ce cas particulier. Un calcul sans difficulté établit la généralité de ce résultat.

Loi de probabilité pour une variable aléatoire continue

Définition

Soit X une variable aléatoire continue, définie sur l'intervalle $[x_a, x_b]$ — où $x_a < x_b$. La probabilité pour que la variable X prenne une valeur quelconque de l'intervalle $[x_1, x_2]$ définit une application F (x_1, x_2) ; les relations **(5)** généralisées s'écrivent dans ce cas :

$$\textbf{(7)} \qquad \begin{cases} Pr\,(x_1 \leqslant X \leqslant x_2) = F\,(x_1, x_2) \\ 0 \leqslant F\,(x_1, x_2) \leqslant 1 \\ F\,(x_a, x_b) = 1 \end{cases}$$

On note que la fonction F (x_1, x_2) dépend des deux limites de l'intervalle. On introduit la *densité de probabilité* avec une seule variable définie par :

$$\textbf{(8)} \qquad f\,(x) = \frac{d}{dx} F\,(x_a, x),$$

d'où :

$$\textbf{(9)} \qquad F\,(x_1, x_2) = \int_{x_1}^{x_2} f\,(x)\,dx$$

Si dx est un intervalle infinitésimal en x, les **relations (7)** s'écrivent, au vu de **(9)** :



▲ ***Figure 4 :***
le problème de l'urne.

$$\textbf{(10)} \qquad \begin{aligned} &Pr\,(x \leqslant X \leqslant x + dx) = f\,(x)\,dx \\ &0 \leqslant \int_{x_1}^{x_2} f\,(x)\,dx \leqslant 1 \\ &\int_{x_a}^{x_b} f\,(x)\,dx = 1. \end{aligned}$$

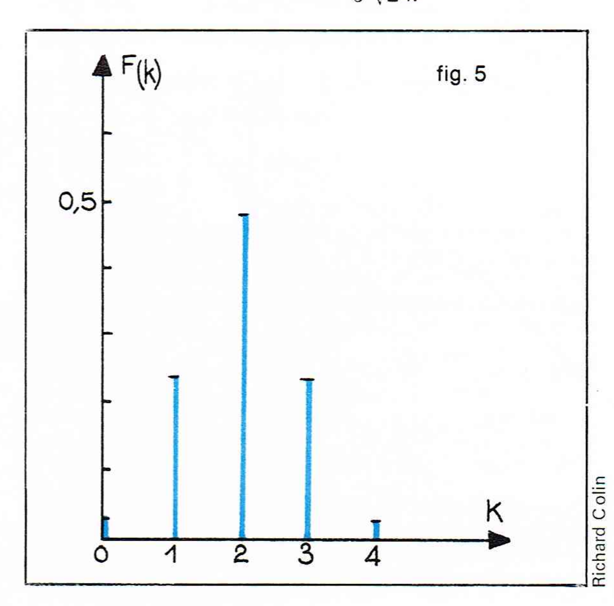
La fonction F (x_a, x) étant monotone et croissante dans l'intervalle (x_a, x_b), $f\,(x)$ est définie positive par **(8)**. Dans les relations ci-dessus nous supposons implicitement $x_1 \leqslant x_2$; dans le cas contraire, nous devrions prendre les valeurs absolues des intégrales.

Les notions de *distributions de probabilité* F (x_a, x) et de *densité de probabilité* $f\,(x)$ se rapportant à une variable aléatoire continue sont d'une importance fondamentale, en particulier pour la physique statistique.

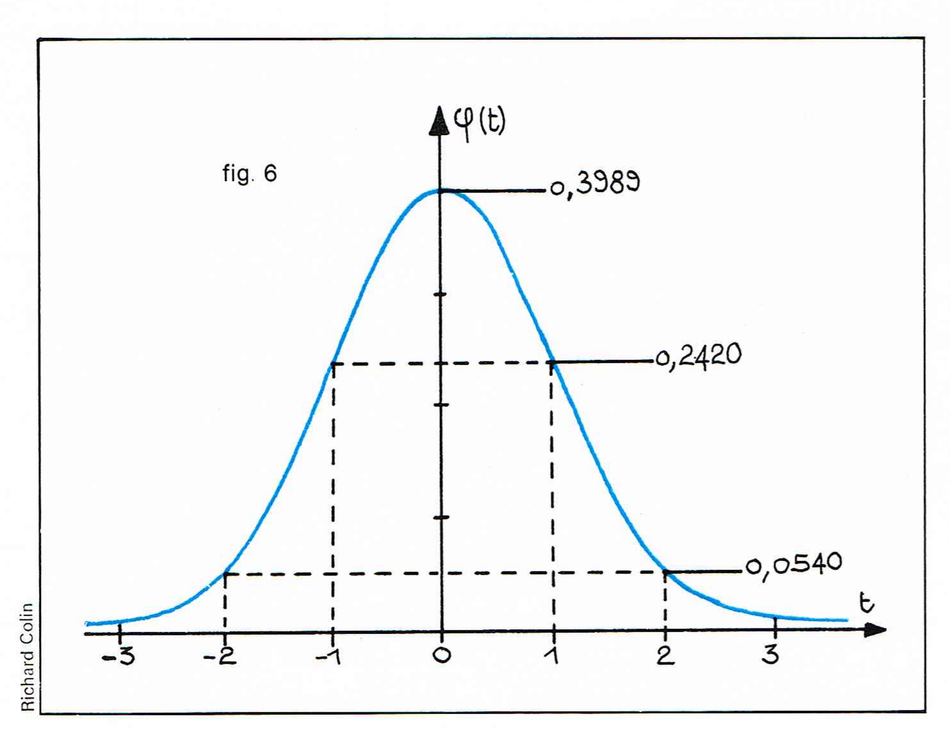
La distribution de Gauss

C'est surtout l'immense champ d'applications de la distribution de Gauss qui justifie la réflexion précédente. Celle-ci, qu'on appelle aussi *distribution normale*, est définie sur l'intervalle $(-\infty, +\infty)$ par la densité associée :

$$\textbf{(11)} \qquad f\,(x) = N\,(x - x_0, \sigma) = \frac{1}{\sigma\sqrt{2\pi}}\,e^{-\frac{(x-x_0)^2}{2\sigma^2}}$$



◄ ***Figure 5 :***
distribution F (k) dans le cas particulier $r = 4$, $n_1 = 5$, $n = 10$;
k peut prendre seulement les valeurs 0, 1, 2, 3 et 4.



▲ ***Figure 6 :*** ***densité associée à la* distribution normale, centrée *et* réduite.**

où σ et x_0 sont deux paramètres, dont le premier est positif. Quand x_0 est nul, on dit que la distribution est *centrée* et quand σ vaut 1, la distribution est *réduite*. On écrit d'habitude la distribution centrée et réduite de la façon suivante :

(12) $$N(t, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} = \varphi(t)$$

Il est clair que le passage de **(11) à (12)** est toujours possible, moyennant un changement de variable aléatoire. La *figure* 6 représente la densité associée à la distribution centrée et réduite. Diverses applications de la distribution de Gauss sont énumérées au long de ce chapitre.

Loi de probabilité se rapportant à plusieurs variables

Cas des variables discrètes

Soit X une variable aléatoire susceptible de prendre les valeurs $x_1, ..., x_n$ et Y une autre variable aléatoire susceptible de prendre les valeurs $y_1, ..., y_n$. A chaque événement élémentaire de l'espace des événements possibles sont alors associées non pas seulement une, mais deux valeurs (x_j, y_k), la première se rapportant à X et la seconde à Y. La probabilité de réalisation des événements correspondants définit une distribution de probabilité $F(x_j, y_k)$ telle que :

$$Pr(X = x_j \text{ et } Y = y_k) = F(x_j, y_k)$$
$$0 \leqslant F(x_j, y_k) \leqslant 1$$
$$\sum_{j,k} F(x_j, y_k) = 1$$

en application des relations **(5)**.

Représentons par $F_1(x_j)$ la distribution de probabilité de la variable aléatoire X, et par $F_2(y_k)$ la distribution de probabilité de la variable aléatoire Y. Par définition $F_1(x_j)$ est la probabilité pour $X = x_j$, indépendamment de la valeur prise par Y. D'où :

(13) $$F_1(x_j) = \sum_k F(x_j, y_k)$$

De même :

(14) $$F_2(y_k) = \sum_j F(x_j, y_k).$$

Les distributions $F_1(x_j)$, $F_2(y_k)$ sont parfois appelées *distributions marginales* de la distribution à deux dimensions $F(x_j, y_k)$. Un cas particulièrement important est celui pour lequel nous avons :

(15) $$F(x_j, y_k) = F_1(x_j) F_2(y_k).$$

Il est facile de voir que dans ce cas les relations **(13)** et **(14)** sont automatiquement satisfaites, car

$$\sum_j F_1(x_j) = \sum_k F_2(y_k) = 1.$$

Quand $F(x_j, y_k)$ se met sous la forme **(15)**, on dit que les variables X et Y sont *indépendantes*. En comparant avec la relation **(3)**, on voit qu'une définition équivalente de l'indépendance des variables X et Y est :

(16) $$Pr\left({}^{Y}\!/_{X}\right) = Pr(Y)$$

et *vice versa*.

Cas des variables continues

L'application du raisonnement précédent — moyennant l'extension au cas de deux variables aléatoires des relations **(10)** — n'offre pas de difficultés particulières. On introduit alors la densité de probabilité $f(x, y)$ pour deux variables aléatoires avec les propriétés :

$$Pr(x \leqslant X \leqslant x + dx \text{ et } y \leqslant Y \leqslant y + dy) = f(x, y)\, dx\, dy$$
$$0 \leqslant \int_{x_1}^{x_2} \int_{y_1}^{y_2} f(x, y)\, dx\, dy \leqslant 1$$
$$\int_{x_a}^{x_b} \int_{y_a}^{y_b} f(x, y)\, dy = 1$$

et :

(17) $$f_1(x) = \int_{y_a}^{y_b} f(x, y)\, dy$$
$$f_2(y) = \int_{x_a}^{x_b} f(x, y)\, dx$$

où $f_1(x)_1$ et $f_2(y)$ sont les densités de probabilité se rapportant à X et Y ; on les appelle aussi *densités marginales de* $f(x, y)$.

De la même façon que précédemment, on dit que X et Y sont indépendantes quand :

(18) $$f(x, y) = f_1(x) f_2(y)$$

Moments d'une densité de probabilité

Définitions

Soit n un entier positif ou nul,

— le moment d'ordre n de la densité de probabilité $f(x)$ est défini par la somme :

$$\mathcal{M}_n = \int_{x_a}^{x_b} x^n f(x)\, dx$$

— le moment d'ordre 0 est appelé *norme* de $f(x)$ et vaut 1, par suite de **(10)** :

$$\mathcal{M}_0 = \int_{x_a}^{x_b} f(x)\, dx = 1$$

— le moment d'ordre 1 est appelé *l'espérance mathématique* de X

(19) $$\mathcal{M}_1 \equiv \langle x \rangle \equiv E(X) = \int_{x_a}^{x_b} x f(x)\, dx$$

De façon plus générale, si $y = y(x)$ est une fonction de x définissant une nouvelle variable aléatoire Y, l'espérance mathématique pour Y est :

(20) $$E(Y) = \int_{x_a}^{x_b} y(x) f(x)\, dx$$

En effet, si $f(x)$ est la probabilité pour que

$$x \leqslant X \leqslant x + dx$$

et $F(y)$ est la probabilité pour que $y \leqslant Y \leqslant y + dy$, nous avons :

$$E(Y) = \int_{y_a}^{y_b} yF(y)\, dy =$$
$$\int_{y_a}^{y_b} yf(x) \left|\frac{dx}{dy}\right| dy = \int_{x_a}^{x_b} y(x) f(x)\, dx$$

Cette relation est immédiate quand $y(x)$ est monotone ; le lecteur peut vérifier qu'elle s'applique aussi en dehors de cette condition. On note que :

(21) $\mathcal{M}_n = E(X^n)$

— la *variance* relative à $f(x)$ est par définition :

(22) $\sigma^2 = E[(X - \langle x \rangle)^2]$

par suite de **(20)**

(23) $\sigma^2 = \int_{x_a}^{x_b} (x - \langle x \rangle)^2 f(x)\, dx$

Soit, en développant l'expression entre parenthèses :

(24) $\sigma^2 = E(X^2) - E(X)^2$

Par suite de **(23)**, σ^2 ne peut être négatif, car l'intégrant comporte seulement des termes définis positifs. La racine positive de la variance, σ, s'appelle *l'écart type* se rapportant à $f(x)$.

Toutes ces définitions s'appliquent, dûment modifiées, aux variables aléatoires discrètes. En particulier, l'espérance mathématique $E(x) = \sum_j x_j F(x_j)$ est la moyenne des x_j pondérés par $F(x_j)$; si les probabilités $F(x_j)$ sont toutes égales, E (X) est la moyenne arithmétique des x_j ; si X est définie comme prenant la valeur 0 quel que soit l'événement $a_j \neq a_i$ et la valeur 1 pour l'événement a_i, E (X) se réduit à la probabilité p_i de l'événement a_i.

Fonction caractéristique

La transformée de Fourier de la densité de probabilité $f(x)$:

(25) $\varphi(k) = \int_{-\infty}^{+\infty} e^{ikx} f(x)\, dx$

existe toujours, car $f(x)$ est une fonction sommable. Elle est appelée *fonction caractéristique* associée à $f(x)$, ou fonction génératrice des moments. Cette dernière dénomination provient du fait qu'en portant le développement en série de Taylor de e^{ikx} autour de $x = 0$ dans **(25)**, on obtient :

$$\varphi(k) = \int \left[1 + ikx - \frac{k^2}{2} x^2 + \dots + \frac{(ik)^n}{n!} x^n + \dots\right] f(x)\, dx$$

(26) $= \mathcal{M}_0 + ik\mathcal{M}_1 - \frac{k^2}{2}\mathcal{M}_2 + \dots + \frac{(ik)^n}{n!}\mathcal{M}_n + \dots$

Réciproquement, la connaissance de $\varphi(k)$ détermine la distribution $f(x)$. La transformation inverse de **(25)** est :

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-ikx} \varphi(k)\, dk$$

Nous avons alors cet important résultat : *la connaissance de tous les moments d'une distribution de probabilité définit complètement cette distribution* (puisqu'elle donne immédiatement la fonction caractéristique).

La distribution G

Appliquons les considérations précédentes à la distribution G définie sur $(0, +\infty)$ par la densité :

(27) $G(\alpha, \beta, x) = Ax^{\alpha} e^{-\beta x}$

où α et β sont deux paramètres réels positifs et A est la constante de normalisation :

(28) $A = \frac{\beta^{\alpha+1}}{\Gamma(\alpha + 1)}$

où $\Gamma(x+1) \equiv \int_0^{\infty} y^x e^{-y} dy \qquad (-1 < x < +\infty)$

Pour trouver la fonction caractéristique de cette distribution, démontrons le lemme suivant.

Lemme : $\forall \alpha$ réel > 0 et $\forall \beta$ (Re $\beta > 0$),

$$\gamma(\alpha, \beta) = \int_0^{\infty} x^{\alpha} e^{-\beta x} dx = \frac{\Gamma(\alpha+1)}{\beta^{\alpha+1}}.$$

Pour β réel, cette relation est la conséquence immédiate du changement de variable $y = \beta x$.

Pour β complexe, on a, en posant $z = \beta x$

$$\gamma(\alpha, \beta) = \frac{1}{\beta^{\alpha+1}} \int_A^B z^{\alpha} e^{-z} dz \qquad (fig. 7)$$

La fonction $f(z) = z^{\alpha} e^{-z}$, $\alpha > 0$ n'a pas de pôles dans la région Re $z > 0$. En sommant sur le contour ABCA :

$$\oint f(z)\, dz = \int_A^B f(z)\, dz + \int_B^C f(z)\, dz + \int_C^A f(z)\, dz = 0.$$

Quand le rayon du cercle BC tend vers l'infini, $f(z)$ tend vers 0 sur tous les points du parcours. Par suite :

$$\int_A^B f(z)\, dz = \int_A^C f(z)\, dz = \int_0^{\infty} z^{\alpha} e^{-z} dz = \Gamma(\alpha + 1)$$

La fonction caractéristique associée à **(27)** est donc :

$$\varphi(k) = A \int_0^{\infty} x^{\alpha} e^{-(\beta - ik)x} dx = \left[\frac{\beta}{\beta - ik}\right]^{\alpha+1}$$

Les premiers moments de la distribution $G(\alpha, \beta, x)$ sont, en application de **(26)** :

$$\mathcal{M}_0 = \varphi(0) = 1$$

(29) $\mathcal{M}_1 = -i\varphi'(0) = \frac{\alpha + 1}{\beta}$

$$\mathcal{M}_2 = -\varphi''(0) = \frac{(\alpha+1)(\alpha+2)}{\beta^2}$$

Il est facile de vérifier que ces valeurs coïncident avec la définition des moments $\left(\mathcal{M}_n = \int_{x_a}^{x_b} x^n f(x)\, dx\right)$.

On note que, pour la distribution G, E (X) ne coïncide pas avec la *valeur plus probable* de x (définie par $f'(x) = 0$),

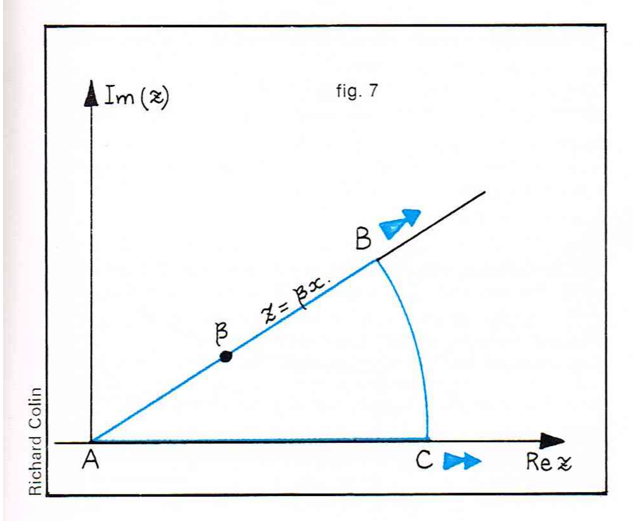
soit $x = \frac{\alpha}{\beta}$

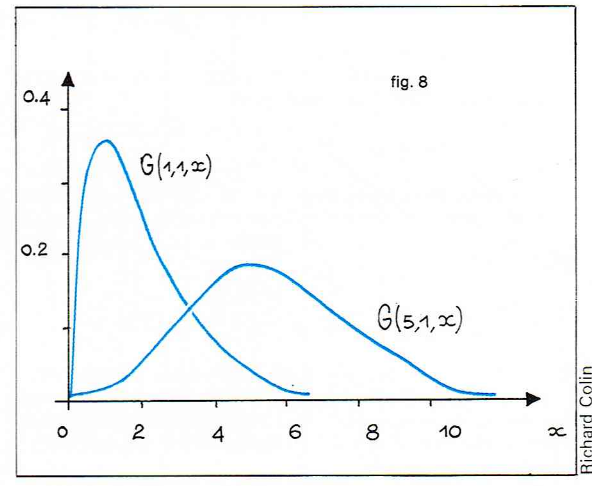
La *figure* 8 représente les distributions $G(1, 1, x)$ et $G(5, 1, x)$.

La distribution de Dirac - Applications

Définition

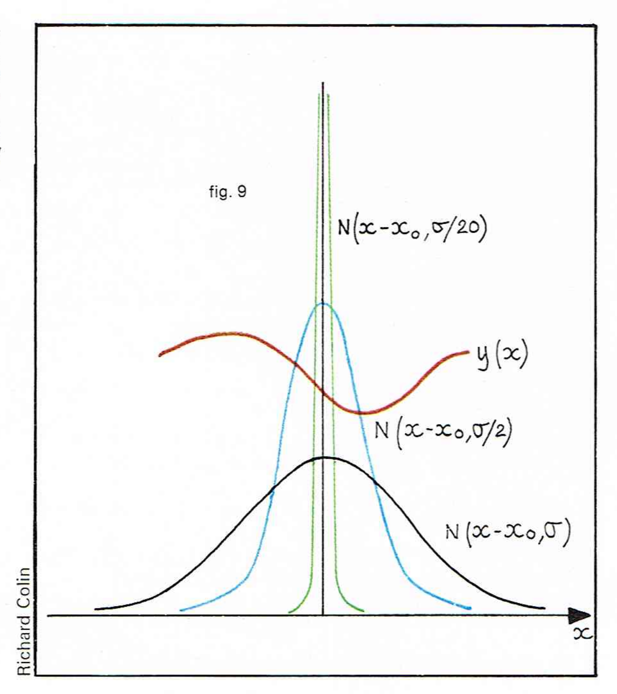
Considérons la « déformation » subie par la densité normale **(11)** quand, x_0 étant fixé, on fait diminuer la valeur de σ. La valeur centrale $N(0, \sigma) = \frac{1}{\sigma\sqrt{2\pi}}$ augmente, tandis que la courbe se rétrécit autour de la valeur





◄ *A gauche, figure 7 : voir développement dans le texte ci-dessus. A droite, figure 8 : distributions G (1, 1, x) et G (5, 1, x).*

▶ *Figure 9 : pour σ très petit, la densité normale demeure presque nulle en tout point, sauf au voisinage de x_0, où elle prend des valeurs très grandes.*



$x = x_0$, de telle sorte que la condition de normalisation reste vérifiée : $\forall \sigma > 0$, $\int_{-\infty}^{+\infty} N(x - x_0, \sigma)\, dx = 1$. Pour σ très petit, la distribution demeure presque nulle en tout point, sauf au voisinage de x_0, où elle prend des valeurs très grandes (*fig.* 9). Considérons maintenant une somme du type : $\mathfrak{I} = \int_{-\infty}^{+\infty} y(x)\, N(x - x_0, \sigma)\, dx$ où $y(x)$ est une fonction « suffisamment » régulière en x_0. Il est clair que, à mesure que σ diminue, l'expression au-dessous du signe somme ne prend de valeurs non négligeables qu'au voisinage immédiat de $x = x_0$. Dans ce voisinage, $y(x)$ reste voisin de la valeur $y(x_0)$, de sorte que :

$$\lim_{\sigma \to 0} \mathfrak{I} = y(x_0) \int_{-\infty}^{+\infty} N(x - x_0, \sigma)\, dx = y(x_0).$$

Nous définirons *la densité δ de Dirac* à travers la propriété fondamentale suivante : $y(x)$ étant une fonction bien définie en x_0,

$$\textbf{(30)} \qquad \int_{-\infty}^{+\infty} y(x)\, \delta(x - x_0)\, dx = y(x_0).$$

Cette densité peut être représentée symboliquement par :

$$\delta(x - x_0) = \lim_{\sigma \to 0} N(x - x_0, \sigma).$$

Cette représentation de la distribution δ est trop étroite pour couvrir toute la généralité de la propriété **(30)**, pour laquelle nous n'avons imposé aucune condition de régularité sur la fonction $y(x)$; par suite, elle ne constitue pas une définition de la densité. Elle suffit cependant dans les applications pratiques que nous ferons.

Propriétés de la densité δ

$\delta(x - x_0)$ étant nulle, quel que soit $x \neq x_0$, la propriété fondamentale **(30)** peut aussi s'écrire :

$$\int_{x_1}^{x_2} y(x)\, \delta(x - x_0)\, dx = \begin{cases} y(x_0) & \text{si } x_0 \in]x_1, x_2[\\ 0 & \text{si } x_0 \notin [x_1, x_2] \end{cases}$$

En effectuant un changement de variable convenable dans la relation ci-dessus, le lecteur vérifiera que, si $g(x)$ est une fonction nulle aux points x_i ($i = 1, n$) et telle que $\forall i\ g'(x_i) \neq 0$, on a :

$$\textbf{(31)} \qquad \delta(g(x)) = \sum_i \frac{1}{|g'(x_i)|}\, \delta(x - x_i)$$

Application aux fonctions de variables aléatoires

Soit $Y = g(X)$ une fonction de la variable aléatoire X définie sur l'intervalle $[x_a, x_b]$ et régie par la densité de probabilité $f(x)$. La densité de probabilité $F(y)$ relative à Y est donnée par la relation :

$$\textbf{(32)} \qquad F(y) = \int_{x_a}^{x_b} f(x)\, \delta(y - g(x))\, dx$$

En effet, les moments de la distribution $F(y)$, qui, comme nous l'avons vu, définissent complètement cette distribution, sont :

$$\mathcal{M}_n = \int_{y_a}^{y_b} y^n F(y)\, dy = \int_{y_a}^{y_b}\int_{x_a}^{x_b} y^n f(x)\, \delta(y - g(x))\, dx dy = \int_{x_a}^{x_b} [g(x)]^n f(x)\, dx$$

ce qui concorde bien avec **(20)** et **(21)**.

La relation **(32)** s'étend immédiatement au cas d'une variable aléatoire Y fonction de différentes variables aléatoires $X_1, X_2, ..., X_n$; si $f(x_1, x_2, ..., x_n)$ est la densité de probabilité (à n dimensions) relative à $X_1 ..., X_n$ et si Y est définie par la fonction $Y = g(x_1, x_2, ..., x_n)$, on a :

$$\textbf{(33)} \qquad F(y) = \int f(x_1, x_2, ..., x_n)\, \delta(y - g(x_1, x_2, ..., x_n))\, dx_1 dx_2 ... dx_n$$

Somme de variables aléatoires

Les sommes de variables aléatoires vérifient les propriétés fondamentales suivantes :

— *Espérance mathématique*

Si $Y = X_1 + X_2 + ... + X_n$, alors

$$\textbf{(34)} \qquad E(Y) = E(X_1) + E(X_2) + ... + E(X_n)$$

— *Variance*

Si $Y = X_1 + X_2 + ... + X_n$ et si les variables $X_1, X_2, ... X_n$ sont indépendantes, alors

$$\textbf{(35)} \qquad \sigma_y^2 = \sigma_{x1}^2 + \sigma_{x2}^2 + ... + \sigma_{xn}^2$$

Les relations **(34)** et **(35)** sont simplement des applications des définitions **(19)** et **(23)** sur la distribution **(33)**, avec $g(x_1, x_2, ..., x_n) = x_1 + x_2 ... + x_n$ [quand $X_1, X_2, ..., X_n$ sont indépendantes, $f(x_1, x_2, ..., x_n)$ est de la forme $f_1(x_1) f_2(x_2) ... f_n(x_n)$ — voir **(18)**].

La distribution binominale et ses approximations

La distribution binominale

La distribution binominale se rapporte aux problèmes très généraux d'*épreuves répétées* (simultanément ou successivement) sur une même variable aléatoire X, régie, à chaque tirage, par la même loi de probabilité $f(x)$.

Définition

Soit $\mathcal{E}$ un espace d'événements possibles, A un événement $\in \mathcal{E}$. Dans un essai isolé, A a la probabilité p de se produire. La distribution binominale donne la probabilité d'obtenir k épreuves favorables, c'est-à-dire, donnant A pour résultat, quand l'épreuve est répétée n fois.

Pour chaque essai, nous avons deux résultats possibles : ou A se produit (probabilité p), ou $\bar{A}$, l'événement complémentaire, se produit (probabilité $q = 1 - p$). Considérons une séquence quelconque de n événements ; par exemple, la séquence $S = A, \bar{A}, A, \bar{A}, \bar{A}$, etc., dans laquelle l'événement A se produit k fois. Cette séquence a pour probabilité :

$$Pr(S) = Pr\{A \cap \bar{A} \cap A \cap \bar{A} \cap \bar{A} ...\} = pqpqq ... = p^k q^{n-k}$$

d'après **(3)** et **(16)**. En effet, c'est la probabilité de l'événement composé « A se produit dans le premier essai et $\bar{A}$ dans le second, etc. », compte tenu du fait que les événements successifs sont indépendants.

Le nombre de séquences telles que S, contenant exactement k fois A, est $\binom{n}{k}$. La réalisation de chacune de ces séquences exclut celle des autres, de sorte que la probabilité pour que l'une ou l'autre de ces séquences se produise est :

$$\textbf{(36)} \qquad P(k) = \binom{n}{k} p^k q^{n-k}$$

en application de **(4)**.

Exemple. Dans le jeu de pile ou face, la probabilité de « pile » est $\frac{1}{2}$. La probabilité d'obtenir 4 fois « pile » dans dans une partie de 5 essais est :

$$P(4) = \binom{5}{4}\left(\frac{1}{2}\right)^5 \simeq 0{,}152.$$

On note que la distribution **(36)** est différente de la distribution hypergéométrique **(6)**. Dans l'exemple du tirage de boules dans une urne, la distribution **(6)** se rapporte au tirage successif de r boules dans une même urne (de sorte que la probabilité d'obtenir une boule blanche varie d'un tirage à l'autre). Au contraire, la distribution **(36)** se rapporte au tirage de n boules dans n urnes identiques, ou encore au tirage de n boules dans une urne, à condition que l'on recompose l'urne initiale après chaque tirage.

Moments

Les moments de la distribution **(36)** se déduisent de la *loi du binôme de Newton :*

$$\sum_{i=0}^{n} \binom{n}{i} a^{n-i} b^i = (a+b)^n$$

qui donne immédiatement

$$\mathcal{M}_0 = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} = (p+q)^n = 1.$$

Le calcul des moments d'ordre supérieur s'effectue de même, en posant $k' = k - 1$. On trouve :

$$\textbf{(37)} \qquad \begin{aligned} E(k) &= np \\ \sigma^2(k) &= npq. \end{aligned}$$

Distribution de Poisson

Définition ; moments

La distribution de Poisson est une approximation de la distribution binominale, valable quand la probabilité p est très petite et le nombre de tirages n très grand, de sorte que le nombre de cas favorables ($\sim np$) reste de l'ordre de quelques unités.

Posons $\lambda = np$; on a :

$$P(k) = \frac{n!}{k!\,(n-k)!} p^k q^{n-k} \simeq \frac{1}{k!} n^k \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^n$$

puisque $n \gg k$. Pour n très grand, nous avons :

$$\operatorname{Log}\left(1 - \frac{\lambda}{n}\right)^n = n \operatorname{Log}\left(1 - \frac{\lambda}{n}\right)$$

$$= n\left(-\frac{\lambda}{n} - \frac{\lambda^2}{2n^2} \cdots\right) \simeq -\lambda$$

et

$$P(k) \simeq \frac{\lambda^k e^{-\lambda}}{k!}.$$

Cette relation définit la *distribution de Poisson*.

Les moments correspondants se calculent aisément, compte tenu de la relation $\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{\lambda}$. On trouve :

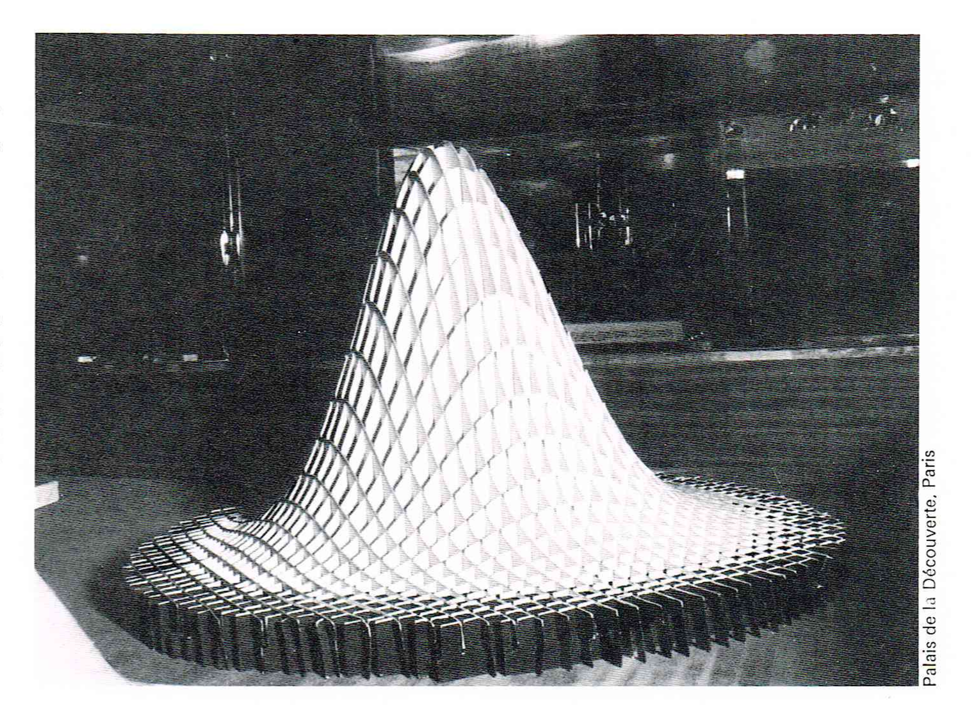
$$\begin{aligned} \mathcal{M}_0 &= 1 \\ E(k) &= \lambda \\ \sigma^2(k) &= \lambda. \end{aligned}$$

Exemple d'application : substances radio-actives

Un exemple courant de l'application de l'approximation de Poisson est la désintégration d'une substance radioactive de grande durée de vie τ. Un échantillon d'une telle substance est composé d'un nombre très grand : $n \sim 10^{23}$ atomes, chacun d'eux ayant une probabilité très petite de se désintégrer pendant la durée de l'observation $\Delta t \ll \tau$.

Ainsi, le nombre moyen de désintégrations observées dans un microgramme de Ra^{226} ($\tau \simeq 5 \cdot 10^{10}$ s) à travers un angle solide de 0,002 4 sr en une seconde, est

$$\lambda = np \simeq 10.$$



▲ ***Surface de Gauss.***

Palais de la Découverte, Paris

La *figure* 10 représente la distribution de probabilité d'observer k désintégrations en une seconde. On note que si la probabilité p n'était pas aussi petite (durée de vie du même ordre de grandeur que le temps de mesure et grand angle solide), l'approximation de Poisson deviendrait injustifiée dans cette application.

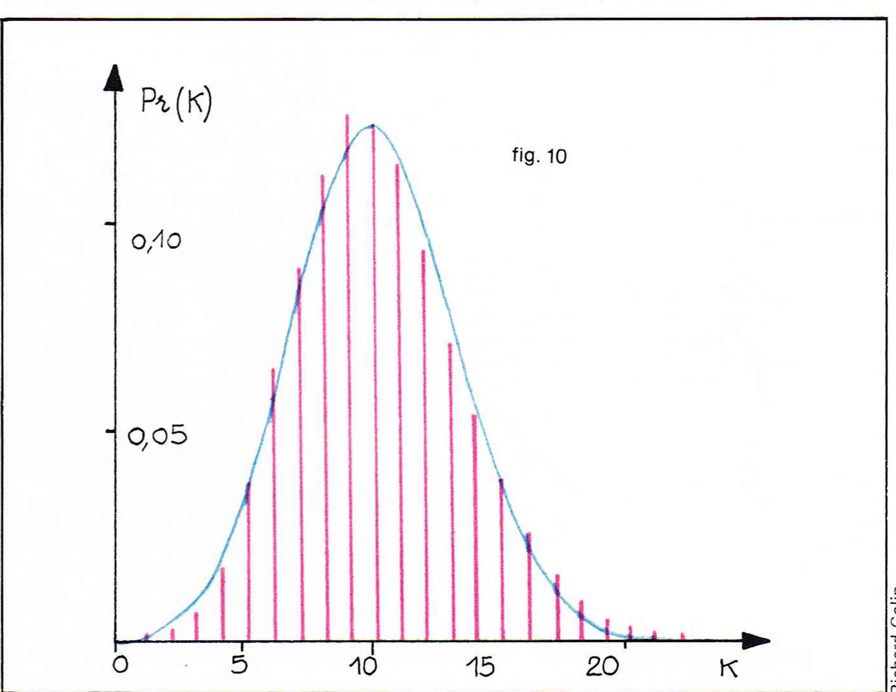
Distribution de Gauss

La distribution de Gauss ou normale définie en **(11)** constitue une autre approximation de la distribution binominale **(36)** *dans la limite où $n \to \infty$, p et q restant fixés*. Dans cette limite, le nombre de succès $\sim np \to \infty$ avec n, de même que le nombre de cas défavorables $\sim nq$. On a :

$$P(k) = \frac{n!}{k!\,(n-k)!} p^k q^{n-k}$$

$$\textbf{(38)} \qquad \xrightarrow[n \to \infty]{} N(k - np, \sqrt{npq})$$

En effet, quand $n \gg 1$, P (k) ne prend de valeurs appréciables que dans le voisinage de E $(k) = np$, et les



▼ ***Figure 10 : distribution en une seconde, k désintégrations d'une substance radio-active subissant en moyenne 10 désintégrations par seconde.***

Richard Colin

factorielles en n, k et $n - k$ tendent vers la fonction de Stirling :

$$z! \simeq \sqrt{2\pi z}\left(\frac{z}{e}\right)^z$$

Utilisant cette expression pour les factorielles présentes dans **(38)** :

$$P(k) \rightarrow \frac{\sqrt{n}}{\sqrt{2\pi k(n-k)}}\left(\frac{np}{k}\right)^k\left(\frac{nq}{n-k}\right)^{n-k}$$

Posant $\xi = k - np$

$$P(k) \simeq \frac{\sqrt{n}}{\sqrt{2\pi(np+\xi)(nq-\xi)}}\left(\frac{np}{np+\xi}\right)^{np+\xi}\left(\frac{nq}{nq-\xi}\right)^{nq-\xi}$$

$$\simeq \frac{\sqrt{n}}{\sqrt{2\pi(np+\xi)(nq-\xi)}} \exp -\left\{(np+\xi)\,\mathrm{Log}\left(1+\frac{\xi}{np}\right) + (nq-\xi)\,\mathrm{Log}\left(1-\frac{\xi}{nq}\right)\right\}$$

Pour $\xi \ll np$ et $\xi \ll nq$, on obtient, en développant les logarithmes et en ne retenant que les termes dominants :

$$P(k) \simeq \frac{1}{\sqrt{2\pi npq}} \exp - \frac{\xi^2}{2npq} = N(k - np, \sqrt{npq})$$

Moments

Les moments de la distribution normale sont :

$$\mathcal{M}_n(\xi) = \int_{-\infty}^{+\infty} \xi^n N(\xi, \sqrt{npq})\, d\xi, \quad \xi = k - np$$

En se servant de l'intégrale définie $\int_0^\infty e^{-t^2}dt = \frac{\sqrt{\pi}}{2}$ et en intégrant par parties, on obtient pour les premiers moments :

$$\mathcal{M}_0 = 1$$
$$E(k) = np$$
$$\sigma^2(k) = npq$$

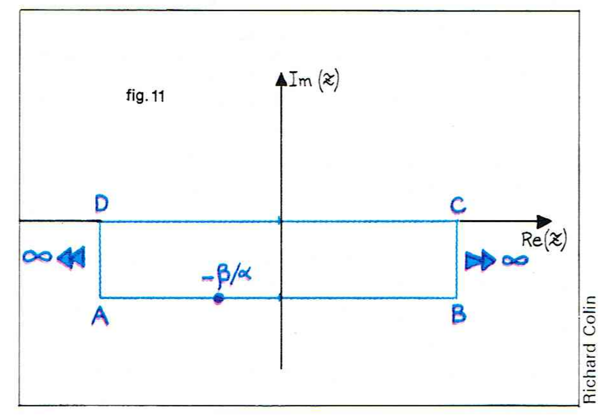
Ces relations sont identiques aux relations **(37)** indiquant que les trois premiers moments de la distribution de Gauss $N(k - np, \sqrt{npq})$ coïncident avec les moments de la distribution binominale. La comparaison des deux distributions montre que l'approximation gaussienne est satisfaisante dans les applications pratiques, dès que $n \gtrsim 10$ (*fig.* 10).

Fonction caractéristique

La fonction caractéristique **(25)** de la distribution normale $\varphi(k) = \int_{-\infty}^{+\infty} e^{ikx} N(x - x_0, \sigma)\, dx$

(39) $$\varphi(k) = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{ikx} e^{-(x-x_0)^2/2\sigma^2}\, dx$$

peut se calculer en utilisant le lemme suivant.



▶ *Figure 11 : voir développement dans le texte ci-contre.*

Lemme

$$\mathfrak{J}(\alpha, \beta) = \int_{-\infty}^{+\infty} e^{-\alpha^2x^2+2\beta x}\, dx,\ \beta \text{ réel ou complexe}$$

(40) $$= \frac{\sqrt{\pi}}{\alpha} e^{\beta^2/\alpha^2}$$

En effet

$$\mathfrak{J}(\alpha, \beta) = e^{\beta^2/\alpha^2}\int_{-\infty}^{+\infty} e^{-\left(\alpha x - \frac{\beta}{\alpha}\right)^2} dx$$

Posant $z = \alpha x - \frac{\beta}{\alpha}$.

$$\mathfrak{J}(\alpha, \beta) = \frac{1}{\alpha} e^{\beta^2/\alpha^2}\int_A^B e^{-z^2}\, dz \qquad (fig.\ 11)$$

La fonction $f(z) = e^{-z^2}$ n'a pas de pôles dans le plan complexe. Intégrant sur le contour ABCD :

$$\oint f(z)\,dz = \int_A^B f(z)\,dz + \int_B^C f(z)\,dz + \int_C^D f(z)\,dz + \int_D^A f(z)\,dz = 0$$

Quand les côtés BC et AD tendent vers l'infini, $f(z)$ tend vers 0 sur tous les points des parcours correspondants. D'où :

$$\int_A^B f(z)\,dz = \int_D^C f(z)\,dz = \int_{-\infty}^{+\infty} e^{-z^2}\,dz, \quad z \text{ réel} = \sqrt{\pi},$$

ce qui complète la démonstration.
Combinant **(39)** et **(40)**, on obtient :

$$\varphi(k) = e^{-\frac{k^2\sigma^2}{2} + ikx_0}$$

Lois des grands nombres

Les lois des grands nombres constituent le noyau de la théorie des probabilités. La première loi, en particulier, assure que la fréquence des événements favorables tend bien vers la probabilité, quand celle-ci existe, conformément à la supposition de base de notre exposé. La seconde loi, d'un autre côté, indique que la distribution de probabilité d'une somme de variables aléatoires tend, presque toujours, vers une distribution normale, à mesure que le nombre de variables de la somme augmente. Comme, dans les cas usuels, la convergence de la distribution vers une distribution normale est très rapide, cette loi permet de comprendre la grande portée de la distribution normale.

Les deux lois des grands nombres peuvent se présenter sous diverses formes, suivant le degré de généralité exigé. Les difficultés des démonstrations augmentent malheureusement avec ce degré de généralité. Pour cette raison, nous nous contenterons d'énoncer ces théorèmes, laissant aux lecteurs intéressés la liberté de chercher dans les livres spécialisés les détails des démonstrations. Les discussions de ces détails appartiennent aux théories mathématiques des probabilités.

Première loi

Soit $\{X_k\}$ une séquence de variables aléatoires indépendantes, chacune d'elles régie par la même distribution de probabilité. Si l'espérance mathématique $E(X) = \mu$ existe, alors, nous avons pour tout $\varepsilon > 0$

$$\lim_{n\to\infty} Pr\left\{\left|\frac{X_1 + \ldots X_n}{n} - \mu\right| > \varepsilon\right\} = 0$$

On notera qu'il n'est pas nécessaire que les autres moments de la distribution de probabilité, et en particulier σ, existent (c'est-à-dire soient finis).

La première loi des grands nombres peut s'appliquer aussi quand les variables aléatoires $X_1, X_2, \ldots, X_n$ ne suivent pas la même distribution de probabilité. Dans ce cas, toutefois, les diverses distributions doivent remplir certaines conditions supplémentaires. Le théorème s'applique, en particulier, si l'une des deux conditions suivantes au moins est satisfaite :

a - toutes les variables X_i sont limitées, c'est-à-dire $\exists$ un nombre A tel que $\forall i\ |X_i| < A$;

b - les variances existent, et S^2 étant la variance de la somme ($S^2 = \Sigma\, \sigma_i^2$), $\frac{S}{n} \rightarrow 0$ quand $n \rightarrow \infty$.

L'espérance mathématique de la somme étant donnée par **(34)**, la limite μ sera dans ce cas la moyenne arithmétique des espérances mathématiques respectives :

$$\mu = \frac{1}{n} \Sigma \mu_i.$$

Appliquons maintenant la première loi des grands nombres aux variables aléatoires X_i qui prennent les valeurs 1 ou 0 selon que le résultat du i-ième essai d'une séquence d'*épreuves répétées* est favorable ou non. $Y = \Sigma X_i$ est le nombre d'événements favorables obtenus au total, et $f = \Sigma \frac{X_i}{n}$ la fréquence. Par ailleurs,

$$E(X) = 0(1-p) + 1(p) = p$$

est la probabilité de succès pour n'importe quel essai. La première loi assure donc que la fréquence tend vers la probabilité, quand le nombre d'épreuves tend vers l'infini.

Deuxième loi (Théorème de la limite centrale)

Soit $\{X_k\}$ une séquence de variables aléatoires indépendantes, chacune d'elles régie par la même distribution de probabilité. Si l'espérance mathématique $E(X) = \mu$ et la variance $E(X^2) - E(X)^2 = \sigma^2$ existent, alors pour tout couple α, β ($\alpha < \beta$) :

$$\lim_{n \to \infty} Pr\left\{ \alpha < \frac{X_1 + \dots X_n - n\mu}{\sigma\sqrt{n}} < \beta \right\} = \Phi(\beta) - \Phi(\alpha)$$

où

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt$$

Pour montrer plus clairement que ce théorème implique que la distribution de la somme $Y = \Sigma X_i$ tend vers une distribution normale, posons :

$$T = \frac{(Y - \bar{Y})}{\sigma\sqrt{n}} \qquad \alpha = t \qquad \beta = t + dt$$

nous obtenons :

$$Pr\{t < T < t + dt\} = N(t, 1)$$

D'où, en se servant de **(32)** :

$$F(y) = \int N(t, 1)\, \delta(y - t\sigma\sqrt{n} - \langle y \rangle)\, dt$$
$$F(y) = N(y - \langle y \rangle, \sigma\sqrt{n})$$

Comme la première loi, la seconde loi des grands nombres peut s'étendre au cas où les variables aléatoires $X_1, \dots, X_n$ ne suivent pas la même distribution de probabilité, pourvu que les diverses distributions satisfassent certaines conditions générales. Elle s'applique en particulier quand la condition **a** définie plus haut est satisfaite et que l'écart type correspondant à la somme, soit S_n, tend vers l'infini quand $n \to \infty$.

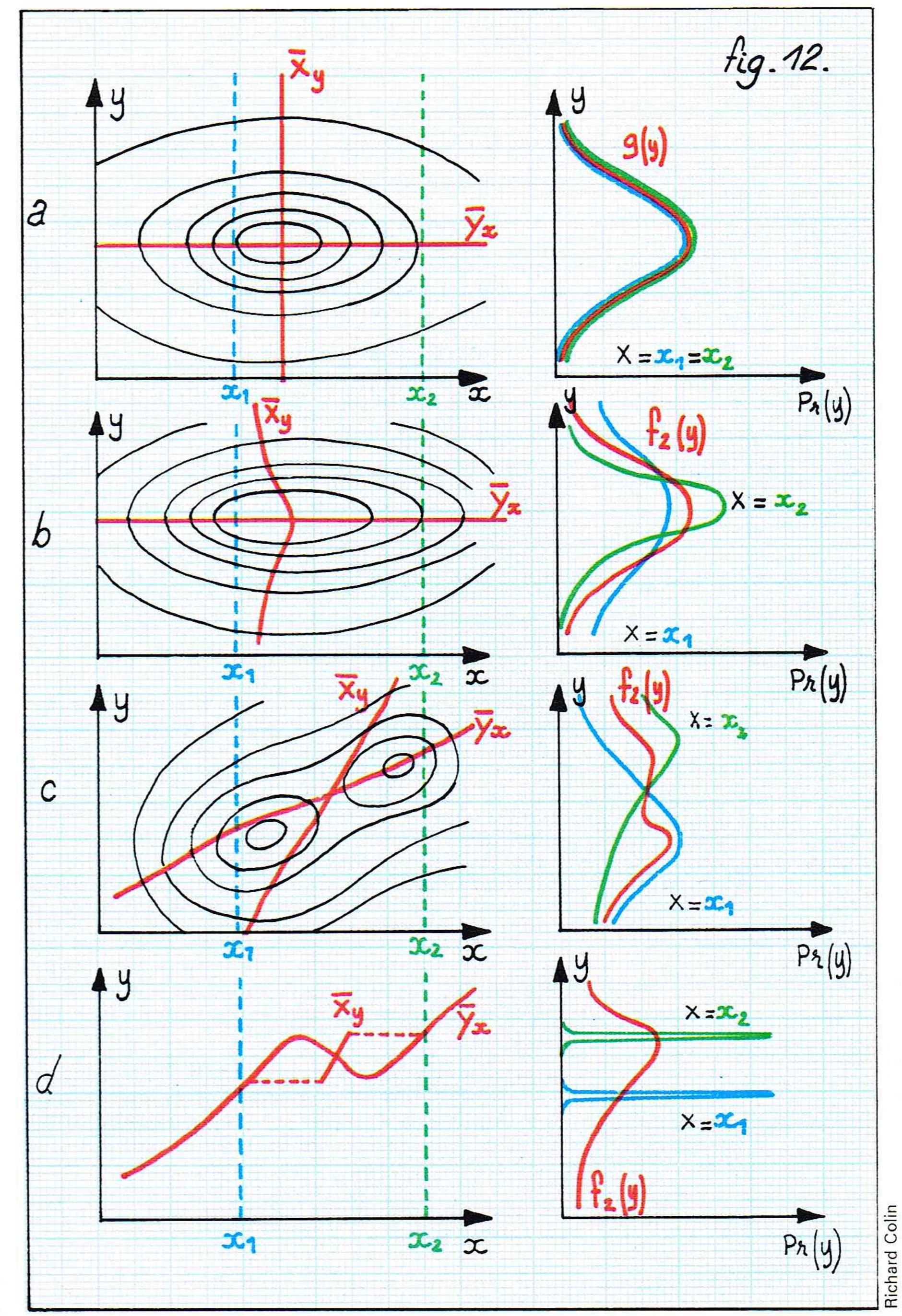
Si X_i prend la valeur 1 ou 0 selon que le résultat du i-ième essai d'une séquence d'épreuves répétées est favorable ou non, nous obtenons le résultat suivant : la limite, pour $n \to \infty$, de la distribution binominale est une distribution normale. Dans ce cas, la seconde loi implique aussi que la fréquence doit tendre vers la probabilité. En effet, à mesure que n augmente, la variance de la fréquence $\frac{Y}{n}$ — soit $\frac{\sigma}{\sqrt{n}}$ — diminue, la distribution se rétrécit autour de l'espérance mathématique, qui est, dans ce cas, la probabilité elle-même.

Corrélations

Par définition, deux événements aléatoires A et B sont dits *indépendants* si la réalisation de l'un quelconque d'entre eux n'influe pas sur la réalisation de l'autre. La probabilité de la réalisation de l'événement composé $A \cap B$ (A et B) est alors simplement le produit des probabilités de réalisation de chacun d'eux.

Si X et Y sont deux variables aléatoires indépendantes et $f_1(x)$ et $f_2(y)$ sont les densités de probabilité associées, la densité de probabilité associée à $X \cap Y$ est

$$f(x, y) = f_1(x)\, f_2(y). \qquad \textbf{(18)}$$



▲ *Figure 12 : quatre exemples de densités de probabilité à deux dimensions pour deux variables aléatoires X et Y diversement liées : a, X et Y sont indépendantes ; b, elles sont liées, bien que la droite de régression $\bar{Y}x$ soit parallèle à l'axe des x ; c, cas général ; d, cas des deux variables liées par la relation Y = Y (X).*

Deux événements aléatoires tels que la réalisation de l'un d'eux modifie les probabilités de réalisation de l'autre sont dits *liés*. La probabilité de réalisation de l'événement composé $A \cap B$ est donnée par **(3)**. La densité de probabilité associée à deux variables aléatoires X et Y liées ne prend pas la forme simple **(18)**.

On dit aussi, souvent, que les variables X et Y sont *corrélées*. Il faut toutefois utiliser ce terme avec prudence, car, comme on le verra, on introduit un coefficient de corrélation entre X et Y qui peut être nul, même pour deux variables liées.

Courbes de régression

Considérons la densité de probabilité à deux dimensions $f(x, y)$ et portons-la sur un diagramme à deux dimensions (x, y), par exemple sous la forme de courbes de niveau (*fig.* 12). La densité de probabilité associée à Y, pour x fixé, est proportionnelle à $f(x, y)$:

$$\textbf{(41)} \qquad \underset{X = x}{g(y)} = f(x, y) \Big/ \int f(x, y)\, dy$$

Si X et Y sont deux variables indépendantes, cette densité est :

$$g_{X=x}(y) = f_1(x) f_2(y) / f_1(x) \int f_2(y)\, dy = f_2(y).$$

Elle reste constamment égale à la distribution marginale $f_2(y)$. En langage imagé, on dira que n'importe quelle coupe effectuée à $x =$ cte reproduit la distribution marginale associée à y. Le fait que les distributions obtenues à $x =$ cte reproduisent toutes la distribution marginale ne fait que traduire directement l'indépendance mutuelle de X et Y.

Cette propriété n'est évidemment plus vraie dans le cas de deux variables aléatoires liées. Considérons la valeur moyenne z de la distribution **(41)**. z est la valeur moyenne des résultats y que l'on peut obtenir sachant que l'on a obtenu $X = x$. Par définition de la moyenne **(19)** :

$$z = \int y g_{X=x}(y)\, dy = \int y f(x, y)\, dy \,/ \int f(x, y)\, dy$$

La courbe décrite par z quand on fait varier x tout au long de l'intervalle (x_a, x_b) de définition de X s'appelle la *courbe de régression de Y en X*.

On peut définir de la même manière la *courbe de régression de X en Y*, définie par l'espérance mathématique de X obtenue à y fixé, quand y varie tout au long de l'intervalle (y_a, y_b) de définition de Y.

Il est évident que, dans le cas où X et Y sont indépendantes, les courbes de régression de Y en X et de X en Y sont des droites parallèles, respectivement à l'axe des x et à l'axe des y. Cependant, le fait que les courbes de régression soient des droites parallèles aux axes n'est pas une condition suffisante pour que X et Y soient indépendantes. En effet, non seulement le moment d'ordre 1, (l'espérance mathématique), mais la distribution de probabilité elle-même de Y à x fixé (ou de X à y fixé) ne doit pas dépendre de x (ou de y). Par exemple, sur la *figure* 12 *b*), la courbe de régression $z = \varphi(x)$ est parallèle à l'axe des x, mais les variables X et Y ne sont manifestement pas indépendantes.

Un cas extrême de liaison entre X et Y est le cas où Y est une fonction univoque de X : $Y = Y(X)$. Dans ce cas, on dit que Y est strictement liée à X. La densité $f(x, y)$ est nulle partout, sauf aux points $y = y(x)$. La courbe de régression $z = \varphi(x)$ se confond avec cette fonction, et les distributions de probabilité à x fixé sont les distributions δ de Dirac : il n'y a aucune dispersion possible des résultats de tirage de Y à x fixé. En particulier, la variance σ_g^2 associée à la **distribution (41)** est nulle.

Coefficient de corrélation de Pearson

La variance σ^2_g de la distribution de Y pour x fixé donne donc une idée de l'intensité avec laquelle les variables X et Y sont liées. La valeur moyenne de cette grandeur

(42) $$\sigma_D^2 = \int \sigma_g^2 f_1(x)\, dx$$

est reliée au coefficient de *corrélation de Pearson* $\eta_{Y/X}$ par la relation :

(43) $$\eta_{Y/X}^2 = 1 - \sigma_D^2 / \sigma_Y^2$$

où σ_Y^2 est la variance de la distribution marginale $f_2(y)$.

$\eta_{Y/X}^2$ est toujours compris entre 0 et 1. Il est nul si $\sigma_z^2 = 0$, ou $z =$ cte. La courbe de régression de Y en X est alors une droite parallèle à l'axe des x, ce qui n'implique pas nécessairement que X et Y soient indépendantes. Il est égal à 1 quand Y est une fonction univoque de X ($\sigma_D^2 = 0$).

L'équation **(43)** ne définit pas le signe du coefficient de corrélation de Pearson $\eta_{Y/X}$. Si $z = \varphi(x)$ est une fonction monotone, on choisit le signe de dérivée de cette fonction ($\eta_{Y/X}$ positif si un accroissement de x favorise un accroissement de y, et négatif dans le cas contraire).

Coefficient de corrélation ρ

Malgré ses avantages, le coefficient de corrélation de Pearson n'est pas la seule mesure possible du degré de liaison entre deux variables aléatoires. On utilise plus fréquemment encore le coefficient ρ, défini par :

(44) $$\rho = \frac{\langle (x - \langle x \rangle)(y - \langle y \rangle) \rangle}{\sigma_X \sigma_Y},$$

qui est le rapport de l'espérance mathématique du produit des variables centrées au produit des variances correspondantes. ρ peut encore s'écrire :

(45) $$\rho = \frac{\langle xy \rangle - \langle x \rangle \langle y \rangle}{\sigma_X \sigma_Y}$$

Il est, en valeur absolue, inférieur ou égal à 1. En effet, si nous appelons Ξ et H les variables centrées correspondant à X et Y prenant les valeurs ξ et η respectivement,

(46) $$\rho^2 = \frac{\langle \xi\eta \rangle^2}{\langle \xi^2 \rangle \langle \eta^2 \rangle}$$

d'après **(44)** et **(23)**.

D'autre part, quel que soit λ, l'expression $(\lambda\xi - \eta)^2$ est $\geqslant 0$. Par suite :

$$\langle (\lambda\xi - \eta)^2 \rangle \geqslant 0$$
$$\lambda^2 \langle \xi^2 \rangle - 2\lambda \langle \xi\eta \rangle + \langle \eta^2 \rangle \geqslant 0 \quad \forall \lambda;$$

le discriminant réduit de cette expression,

$$\langle \xi\eta \rangle^2 - \langle \xi^2 \rangle \langle \eta^2 \rangle,$$

est donc toujours négatif ou nul *(inégalité de Schwartz)*. Comparant avec **(46)** :

(47) $$\rho^2 \leqslant 1$$

La valeur extrême de ρ, $|\rho| = 1$, est obtenue pour $X - \langle X \rangle = a(Y - \langle Y \rangle)$, c'est-à-dire quand il existe une dépendance linéaire entre les deux variables.

Le coefficient de corrélation de deux variables aléatoires indépendantes est nul ; écrivant ρ conformément à **(45)**, on a, dans ce cas : $\langle xy \rangle = \int xy f_1(x) f_2(y)\, dxdy$
$= \int x f_1(x)\, dx \int y f_2(y)\, dy = \langle x \rangle \langle y \rangle$

A l'opposé, le fait que $\rho = 0$ ne prouve pas que X et Y soient indépendantes.

Loi normale à deux variables

Un cas particulièrement important de distribution à deux dimensions est le cas où la densité de probabilité de Y, à x fixé, est normale [c'est-à-dire est donnée par **(11)**], de même que la loi de probabilité de X à y fixé.

La densité à deux dimensions $f(x, y)$ centrée, définie pour x et $y \in (-\infty, +\infty)$, peut alors se mettre sous la forme :

(48) $$f(x, y) = \frac{\sqrt{\delta}}{2\pi} \exp\left\{ -\frac{1}{2}(ax^2 - 2bxy + cy^2) \right\}$$

où $\delta = ac - b^2$.

La densité marginale de x est :

$$f_1(x) = \frac{\sqrt{\delta}}{2\pi} e^{-\frac{ax^2}{2}} \int_{-\infty}^{+\infty} \exp -\frac{1}{2}(-2bxy + cy^2)\, dy.$$

Le changement de variable $t = y - xb/c$ donne :

(49) $$f_1(x) = N\left(x, \sqrt{\frac{c}{\delta}}\right).$$

De même, la densité marginale de y est :

(50) $$f_2(y) = N\left(y, \sqrt{\frac{a}{\delta}}\right).$$

En utilisant les mêmes notations, la densité de probabilité $g_{X=x}(y)$ de la variable Y à x fixé est :

(51) $$g_{X=x}(y) = N\left(y - \frac{xb}{c}, \frac{1}{\sqrt{c}}\right);$$

tandis que la densité de probabilité $h_{Y=y}(x)$ de la variable Y à y fixé est :

(52) $$h_{Y=y}(x) = N\left(x - \frac{yb}{a}, \frac{1}{\sqrt{a}}\right).$$

La densité **(48)** a donc la propriété remarquable suivante : aussi bien les densités conditionnelles $g(y)$ et $h(x)$ que les densités marginales $f_1(x)$ et $f_2(y)$ sont normales (mais avec des largeurs différentes).

Les équations **(51)**, **(52)** montrent en outre que les courbes de régression sont des droites passant par l'origine. La droite de régression de Y en X a pour pente $\frac{b}{c}$ et celle de X en Y a pour pente $\frac{a}{b}$.

Corrélation

La valeur moyenne du produit XY est :

$$\langle xy \rangle = \iint xy\, f(x, y)\, dx\, dy$$

$$= \frac{\sqrt{\delta}}{2\pi} \int x e^{-\frac{ax^2}{2}} dx \int y \exp -\frac{1}{2}(cy^2 - 2\,bxy)\, dy$$

En utilisant le changement de variable $t = y - \frac{xb}{c}$ et en intégrant par parties l'intégrale sur y, puis celle sur x, on trouve :

(53) $$\langle xy \rangle = \frac{b}{\delta}$$

d'où :

$$\rho = \frac{b}{\sqrt{ac}}$$

où nous avons utilisé **(45)**, **(49)** et **(50)**.

Le coefficient de corrélation de Pearson est donné par **(43)** :

$$\eta^2_{Y/X} = 1 - \sigma^2_D / \sigma^2_Y$$

où $\sigma^2_D = \frac{1}{c}$ et $\sigma^2_Y = \frac{a}{\delta}$ par **(42)**, **(52)** et **(50)** :

$$\eta^2_{Y/X} = \frac{b^2}{ac} = \rho^2.$$

On trouverait de même : $\eta^2_{X/Y} = \frac{b^2}{ac} = \rho^2.$

Pour la distribution normale à deux dimensions, les coefficients de corrélations de Pearson et le coefficient ρ prennent la même valeur absolue. La convention de signe donnée plus haut pour le coefficient $\eta_{Y/X}$ assure qu'ils ont également le même signe.

Réécrivons la distribution **(48)** en fonction des paramètres ρ, σ_X et σ_Y.

On trouve :

(54) $$f(x, y) = \frac{1}{2\pi\, \sigma_X \sigma_Y \sqrt{1 - \rho^2}} \exp -\frac{1}{2}\left\{\frac{1}{1-\rho^2}\left(\frac{x^2}{\sigma^2_X} - 2\rho \frac{xy}{\sigma_X \sigma_Y} + \frac{y^2}{\sigma^2_Y}\right)\right\}$$

Fonction caractéristique

La fonction caractéristique de la distribution **(48)** :

$$\varphi(u, v) = \int\int_{-\infty}^{+\infty} e^{i(ux + vy)} f(x, y)\, dx\, dy$$

se calcule aisément en utilisant le lemme **(40)**. On obtient :

$$\varphi(u, v) = e^{\frac{1}{2\delta}(cu^2 + 2\,buv + av^2)}$$

Il est remarquable que cette relation peut encore s'écrire, en utilisant **(49)**, **(50)** et **(53)** :

(55) $$\varphi(u, v) = \exp -\frac{1}{2}(\langle x^2 \rangle u^2 + 2\langle xy \rangle uv + \langle y^2 \rangle v^2)$$

Notons enfin que la fonction caractéristique $\varphi_1(u)$ associée à la densité marginale $f_1(x)$ est, par définition **(17)** de celle-ci :

(56) $$\varphi_1(u) = \int e^{iux} dx \int f(x, y)\, dy = \varphi(u, o).$$

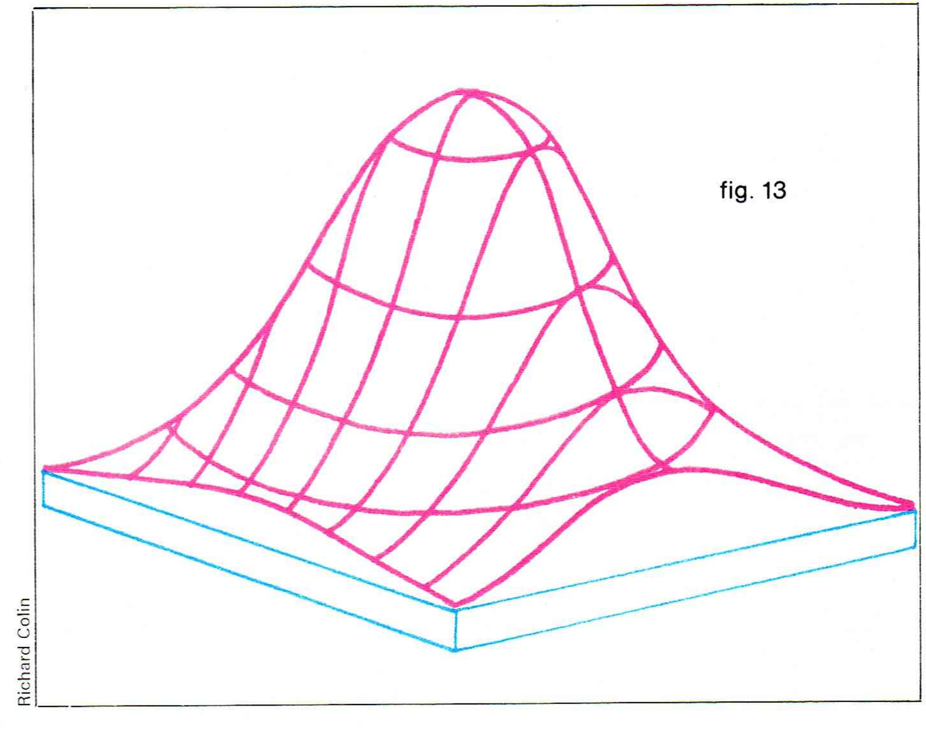
Distribution de deux variables gaussiennes indépendantes

Pour $\rho = 0$ (ou $b = 0$), la densité **(54)** devient simplement :

(57) $$f(x, y) = \frac{1}{2\pi\, \sigma_X \sigma_Y} \exp -\frac{1}{2}\left(\frac{x^2}{\sigma^2_X} + \frac{y^2}{\sigma^2_Y}\right) = N(x, \sigma_X) \times N(y, \sigma_Y).$$

Par conséquent, $f(x, y)$ est dans ce cas la densité de probabilité correspondant à deux variables gaussiennes X et Y indépendantes. Nous avons donc l'importante propriété suivante : pour un couple de variables gaussiennes, la nullité du coefficient de corrélation est équivalente à l'indépendance de ces variables.

La *figure* 13 présente, en perspective, la densité de distribution pour deux variables indépendantes normales, centrées et réduites ($\sigma_X = \sigma_Y = 1$).



▲ ***Figure 13 : représentation, en perspective, de la densité de distribution pour deux variables indépendantes normales, centrées et réduites ($\sigma_X = \sigma_Y = 1$).***

Réduction en facteurs

La transformation linéaire $X' = \alpha X + \beta$ laisse invariant le coefficient de corrélation ρ. En effet,

$$\langle X'Y \rangle = \alpha \langle XY \rangle + \beta \langle Y \rangle, \langle X' \rangle \langle Y \rangle = \alpha \langle X \rangle \langle Y \rangle + \beta \langle Y \rangle \quad \text{et} \quad \sigma'_X = \alpha\sigma_X.$$

Par contre, la transformation plus générale :

(58) $$\begin{cases} X' = \alpha X + \beta Y \\ Y' = \alpha' X + \beta' Y \end{cases}$$

modifie le coefficient de corrélation. Si nous supposons pour simplifier que X et Y sont des variables indépendantes, centrées et réduites, c'est-à-dire telles que $\langle X \rangle = \langle Y \rangle = 0$, $\langle X^2 \rangle = \langle Y^2 \rangle = 1$, on obtient :

(59) $$\begin{aligned} \langle X'Y' \rangle &= \alpha\alpha' + \beta\beta' \\ \sigma^2_{X'} &= \alpha^2 + \beta^2 \\ \sigma^2_{Y'} &= \alpha'^2 + \beta'^2. \end{aligned}$$

D'où :

$$\rho' = \frac{\alpha\alpha' + \beta\beta'}{\sqrt{\alpha^2 + \beta^2}\sqrt{\alpha'^2 + \beta'^2}}$$

Bien entendu, la transformation inverse est possible. Partant de deux variables X' et Y' normales et centrées (ce qui est toujours possible moyennant un changement de variable trivial), d'écarts quadratiques σ'_X et σ'_Y et de corrélation ρ', on peut trouver (d'une infinité de manières) des nombres α, β, α', β' vérifiant **(59)**.

La transformation inverse de **(58)** :

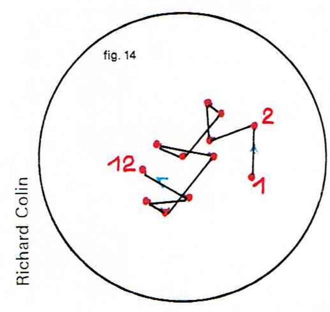
(60) $$\begin{cases} X = (\beta' X' - \beta Y')\,(\alpha\beta' - \alpha'\beta) \\ Y = (-\alpha' X' + \alpha Y')\,(\alpha\beta' - \alpha'\beta) \end{cases}$$

définit deux variables aléatoires X et Y indépendantes, normales, centrées et réduites, dont la densité de probabilité :

$$f(x, y) = N(x, 1)\, N(y, 1)$$

est celle qui est représentée sur la *figure* 13.

Cette transformation s'appelle la *réduction en facteurs* des variables aléatoires gaussiennes X et Y. En statistique, il est extrêmement important, à partir d'un ensemble de variables aléatoires gaussiennes corrélées deux à deux, d'obtenir un ensemble en nombre égal (ou inférieur, si



Richard Colin

▲ *Figure 14 :* **un exemple de processus markovien :** ***les déplacements aléatoires d'une particule sous l'effet des chocs aléatoires des molécules d'un liquide.***

certaines des variables définies ne sont pas linéairement indépendantes) de variables gaussiennes indépendantes, qui sont évidemment plus aptes à caractériser le problème statistique posé. Cette recherche générale des facteurs constitue alors *l'analyse factorielle* (voir *Analyse des données*).

Variables gaussiennes à N dimensions

La forme des équations **(55)** et **(58)** suggère l'avantage qu'il y a, lorsqu'on traite le problème de variables gaussiennes corrélées, de traiter l'ensemble des variables aléatoires sous forme matricielle. La généralisation des résultats obtenus dans le cas de deux variables s'exprime alors ainsi :

a - Généralisation de **(55).**

Soit $\underline{t} = \begin{vmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{vmatrix}$ le vecteur formé par les variables $t_1, ..., t_N$ associées dans la transformation de Fourier aux variables aléatoires $\underline{X} = \begin{vmatrix} X_1 \\ \vdots \\ X_N \end{vmatrix}$

La fonction caractéristique associée à la densité de probabilités de N variables gaussiennes *centrées* est :

$$\varphi(\underline{t}) = e^{-\frac{1}{2}\underline{t}^T \underline{V} \underline{t}}$$

où T symbolise la transposition et où $\underline{V}$ est la matrice des variances de dimension N × N.

$$\textbf{(61)} \qquad \underline{V} = \begin{Vmatrix} \langle X_1^2 \rangle & \langle X_1X_2 \rangle & ... & \langle X_1X_N \rangle \\ \langle X_1X_2 \rangle & \langle X_2^2 \rangle & ... & \langle X_2X_N \rangle \\ \cdots & \cdots & \cdots & \cdots \\ \langle X_1X_N \rangle & \langle X_2X_N \rangle & ... & \langle X_N^2 \rangle \end{Vmatrix}$$

Nous dirons que les variables aléatoires $\underline{X}$ suivent une distribution normale centrée N ($\underline{X}$, $\underline{V}$).

b - Généralisation de **(56).**

Chaque sous-ensemble $\underline{Y} = \begin{vmatrix} X_i \\ \vdots \\ X_k \end{vmatrix}$ à p dimensions (p < N) de X suit une loi N ($\underline{Y}$, $\underline{W}$), où la matrice des variances $\underline{W}$ est formée à partir de la matrice $\underline{V}$ en ne retenant que les lignes et les colonnes correspondant à (X_i, ..., X_k). En d'autres termes, les lois marginales à p dimensions sont normales.

c - Généralisation de **(57).**

Séparons les $\underline{X}$ variables en deux sous-ensembles $\underline{Y}$ et $\underline{Z}$ de dimensions p et q respectivement. Une condition nécessaire et suffisante pour que $\underline{Y}$ et $\underline{Z}$ soient indépendantes est que E ($\underline{Y}\underline{Z}^T$) = 0. En d'autres termes, il faut et il suffit pour cela que toutes les valeurs moyennes du type $\langle Y_k Z_l \rangle$, ($1 \leqslant k \leqslant p$ et $1 \leqslant l \leqslant q$) soient nulles.

d - Généralisation de **(58)** et **(60).**

Soit A une matrice de dimension $k \times$ N. La transformation

$$\textbf{(62)} \qquad \underline{Y} = \underline{A}\underline{X}$$

définit un vecteur aléatoire $\underline{Y}$ qui suit une loi N ($\underline{Y}$, $\underline{W}$), où $\underline{W} = \underline{A}\ \underline{V}\ \underline{A}^T$. Si $\underline{V}$ est régulière, on peut trouver une matrice $\underline{A}$ telle que $\underline{W}$ = 1 (matrice unité), de telle sorte que $\underline{Y}$ suive une distribution N ($\underline{Y}$, 1) normale, centrée et réduite, à k dimensions.

Processus et fonctions stochastiques

Chaînes de Markov

Considérons un espace $\mathcal{E}$ d'événements aléatoires a_i. Nous avons étudié le problème du tirage répété des événements de l'espace $\mathcal{E}$, quand la probabilité affectée à chaque événement a_i ne dépend que de cet événement et reste la même tout au long de la série d'épreuves.

Supposons maintenant que la probabilité affectée à l'événement a_i ne dépende pas seulement de cet événement, mais encore des événements survenus au cours des tirages précédents. On dit que les épreuves successives forment une chaîne. Un cas particulièrement fréquent est celui où les probabilités en question ne dépendent pas des résultats de *tous* les essais précédents, mais seulement du dernier d'entre eux. On dit, dans ce cas, que la chaîne des événements $\{a_i\}$ successifs est une *chaîne de Markov*.

Dans un processus markovien, on a donc à considérer les probabilités $P_{ij} = P_r(a_j/a_i)$ pour que l'événement a_j se produise au n-ième essai, sachant que l'événement a_i s'est produit à l'essai précédent. Un exemple particulièrement remarquable de processus markovien est donné par le mouvement brownien des petites particules colloïdales en suspension dans un liquide, quand on les observe au microscope. Sous l'effet des chocs aléatoires des molécules du liquide, la particule subit des déplacements aléatoires, sa position à l'instant de la n-ième observation dépendant de sa position à la (n — 1)-ième observation (*fig.* 14).

Au lieu de parler de l'espace $\mathcal{E}$ des événements possibles a_i (i = 1, N), on parle souvent de l'espace $\mathcal{E}$ des N états possibles. Le passage de l'état a_i à l'état a_j est alors appelé *transition* $a_i \rightarrow a_j$. Dans le mouvement brownien, il y a une infinité d'états possibles (les positions de la particule colloïdale) et la transition est le passage d'une position à l'autre.

Les probabilités de transition P_{ij} forment une *matrice stochastique* $\underline{P}$. Cette dénomination est liée à la propriété suivante :

$$\sum_{j=1}^{N} P_{ij} = \sum_{j=1}^{N} P_r(a_j/a_i) = 1$$

(la somme des éléments de la matrice le long d'une même ligne est égale à 1), car au cours d'une transition, le système ne peut passer que de l'état i à l'un des N états possibles.

La probabilité pour que le système passe de l'état i à un état j en deux étapes est la somme sur k des transitions $a_i \rightarrow a_k$, $a_k \rightarrow a_j$:

$$P_{ij}^{(2)} = \sum_{k=1}^{N} P_{ik} P_{kj}$$

c'est-à-dire $\qquad \underline{P}^{(2)} = \underline{P}^2$

Par suite : $\qquad \underline{P}^{(n)} = \underline{P}^n$

et si on appelle $\underline{\pi} = \|\pi_1 ... \pi_N\|$ la matrice ligne représentant les probabilités pour que le système se trouve, à l'instant initial, dans l'état (a_1, ..., a_N), la probabilité pour que, après n étapes, les systèmes se trouvent dans l'état j est donnée par :

$$P_j^{(n)} = \sum_i \pi_i P_{ij}^{(n)}$$

soit : $\qquad \underline{P}^{(n)} = \underline{\pi}\, \underline{P}^n$

Une chaîne pour laquelle, à partir de n'importe quel état, on peut atteindre n'importe lequel des autres états (après un nombre variable d'étapes) est appelée *chaîne ergodique*. Mathématiquement, une chaîne ergodique est définie par la propriété suivante : pour tout couple d'états initiaux et finals i et k, la matrice stochastique est telle qu'il existe une puissance $\underline{P}^n$ de $\underline{P}$ telle que

$$(P^n)_{ik} \neq 0.$$

S'il existe une puissance n de $\underline{P}$, soit $\underline{P}^n$, telle que tous les éléments $(P^n)_{ik}$ sont différents de zéro, la chaîne est appelée *régulière*. Dans une chaîne régulière, tous les états peuvent être atteints à partir de n'importe quel état initial, en n étapes exactement.

Équilibre statistique

Les chaînes ergodiques et les chaînes régulières possèdent les importantes propriétés suivantes :

a - pour une chaîne ergodique,

$$\lim_{n \to \infty} \frac{P + P^2 + ... P^n}{n} = A$$

où A est une matrice stochastique dont toutes les lignes sont identiques. La matrice ligne $\underline{\alpha} = \|a_1, a_2 ... a_n\|$ qui les compose est telle que $\underline{\alpha}\underline{P} = \underline{\alpha}$;

b - pour une chaîne régulière :

(63) $$\lim_{n \to \infty} P^n = A$$

où A est une matrice stochastique dont toutes les lignes sont identiques. La matrice ligne $\alpha = \|a_1, a_2 \dots a_n\|$ qui les compose est telle que $\alpha P = \alpha$.

En d'autres termes, pour une chaîne régulière, il arrive un moment où l'addition d'étapes supplémentaires ne modifie plus sensiblement les probabilités pour que le système se retrouve dans tel ou tel état, et ces probabilités sont les mêmes quel que soit l'état initial : un équilibre statistique s'établit entre les diverses transitions possibles.

Nous ne démontrerons pas ces propriétés dans le cas général, mais nous pouvons aisément vérifier la seconde dans le cas simple où N = 2. Dans ce cas, la matrice stochastique s'écrit :

$$P = \begin{vmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{vmatrix}$$

La probabilité $p_{11}^{(n)}$ d'obtenir l'état (1) à partir du même état, après n étapes, est :

$$p_{11}^{(n)} = p_{11}^{(n-1)} p_{11} + p_{12}^{(n-1)} p_{21}.$$

De même :

$$p_{22}^{(n)} = p_{21}^{(n-1)} p_{12} + p_{22}^{(n-1)} p_{22}.$$

Utilisons le fait que :

$$p_{12}^{(n-1)} = 1 - p_{11}^{(n-1)}$$

et

$$p_{21} = 1 - p_{22};$$

$$p_{11}^{(n)} = p_{11}^{(n-1)} p_{11} + (1 - p_{11}^{(n-1)})(1 - p_{22})$$

$$= p_{11}^{(n-1)} (p_{11} + p_{22} - 1) + (1 - p_{22}).$$

Posons :

$$q = (p_{11} + p_{22} - 1)$$
$$r = 1 - p_{11}$$
$$s = 1 - p_{22}.$$

On obtient : $$p_{11}^{(n)} = p_{11}^{(n-1)} q + s.$$

De même : $$p_{22}^{(n)} = p_{22}^{(n-1)} q + r.$$

Par récurrence, on obtient :

(64) $$p_{11}^{(n)} = \frac{r}{1-q} q^n + \frac{s}{1-q}$$

(65) $$p_{22}^{(n)} = \frac{s}{1-q} q^n + \frac{r}{1-q}$$

qui donne bien

$$p_{11}^{(0)} = p_{22}^{(0)} = 1; \quad p_{11}^{(1)} = p_{11}, p_{22}^{(1)} = p_{22}.$$

Les autres termes de P^n sont :

(66) $$p_{12}^{(n)} = 1 - p_{11}^{(n)}, \quad p_{21}^{(n)} = 1 - p_{22}^{(n)}.$$

Si la chaîne à deux états est régulière, $|q| < 1$. En effet, $|q| = 1$ impliquerait, soit $p_{11} = p_{22} = 1$ (les états (1) et (2) restent toujours identiques à eux-mêmes), soit $p_{11} = p_{22} = 0$ (les états (1) et (2) se changent toujours alternativement l'un dans l'autre). On a donc $\lim_{n \to \infty} q^n = 0$.

Par suite :

$$\lim_{n \to \infty} P^{(n)} = A = \begin{vmatrix} \dfrac{s}{1-q} & \dfrac{r}{1-q} \\ \dfrac{s}{1-q} & \dfrac{r}{1-q} \end{vmatrix}$$

où nous avons utilisé **(64)**, **(65)** et **(66)**. La matrice A a bien les propriétés indiquées en **(63)**.

Applications

Une application particulièrement simple des chaînes de Markov est constituée par le problème du cheminement aléatoire à une dimension. Supposons que la puce Pépita puisse sauter le long d'une ligne, de la position 1 à la position N (*fig.* 15). Quand elle se trouve dans une position $k \neq 1 \neq$ N, elle a une probabilité p de sauter vers la position $k + 1$, et une probabilité q de sauter vers la position $k - 1$. Les positions extrêmes 1 et N peuvent être des cages ($p_{11} = p_{NN} = 1$, auquel cas les positions 1 et N sont dites *absorbantes*), ou des ressorts qui renvoient Pépita vers la position adjacente ($p_{12} = p_{N, N-1} = 1$, auquel cas les positions 1 et N sont dites *réflexives*).

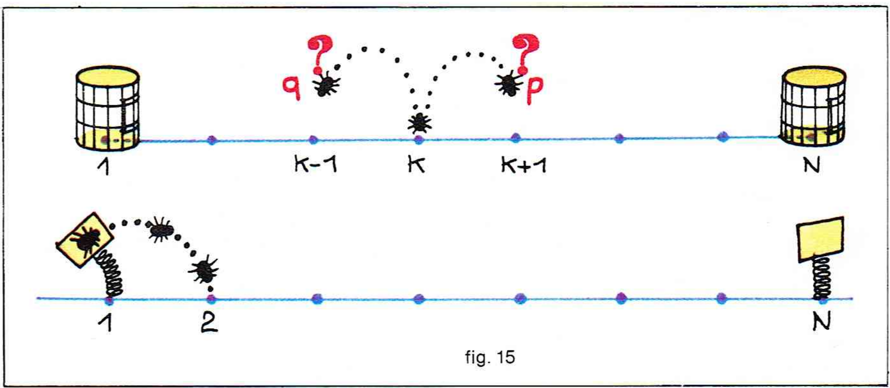


fig. 15

Richard Colin

▲ ***Figure 15 :*** ***chaînes dont les états extrêmes sont absorbants ou réflexifs (cas de la puce Pépita).***

Les matrices stochastiques correspondant à ces deux chaînes sont :

$$P = \begin{vmatrix} 1 & 0 & 0 & \dots & 0 \\ q & 0 & p & \dots & 0 \\ 0 & q & 0 & \dots & 0 \\ \dots & & & & \\ 0 & \dots & q & 0 & p \\ 0 & \dots & 0 & 0 & 1 \end{vmatrix} \qquad P = \begin{vmatrix} 0 & 1 & 0 & \dots & 0 \\ q & 0 & p & \dots & 0 \\ 0 & q & 0 & \dots & 0 \\ \dots & & & & \\ 0 & \dots & q & 0 & p \\ 0 & \dots & 0 & 1 & 0 \end{vmatrix}$$

respectivement. La probabilité pour que, dans le premier de ces exemples, la puce Pépita partie de la position i se retrouve après n sauts dans la cage de la position 1 est $(P^n)_{i1}$. La probabilité pour qu'elle s'y retrouve enfermée après un nombre quelconque de sauts est r_i, que nous pouvons calculer de la façon suivante : au premier saut, la puce a la probabilité p de se retrouver en $i + 1$ et q de se retrouver en $i - 1$. Par conséquent :

(67) $$r_i = p r_{i+1} + q r_{i-1}$$

valable pour $2 \leqslant i \leqslant N - 1$, à condition de poser

(68) $$r_1 = 1, \quad r_N = 0.$$

Pour résoudre cette équation, posons $r_i = x^i$. En substituant r_i dans **(67)**, on obtient : $px^2 - x + q = 0$, qui admet deux solutions $\left(\text{sauf si } p = q = \frac{1}{2}\right)$ $x = 1$ et $x = \frac{q}{p}$. Comme **(67)** est une équation linéaire aux différences finies, la solution générale est :

$$r_i = A + B \left(\frac{q}{p}\right)^i.$$

Les conditions initiales **(68)** fixent A et B. Finalement :

(69) $$r_i = \frac{\left(\dfrac{q}{p}\right)^{N-1} - \left(\dfrac{q}{p}\right)^{i-1}}{\left(\dfrac{q}{p}\right)^{N-1} - 1}$$

Dans le cas où $p = q = \frac{1}{2}$, l'équation **(67)** s'écrit :

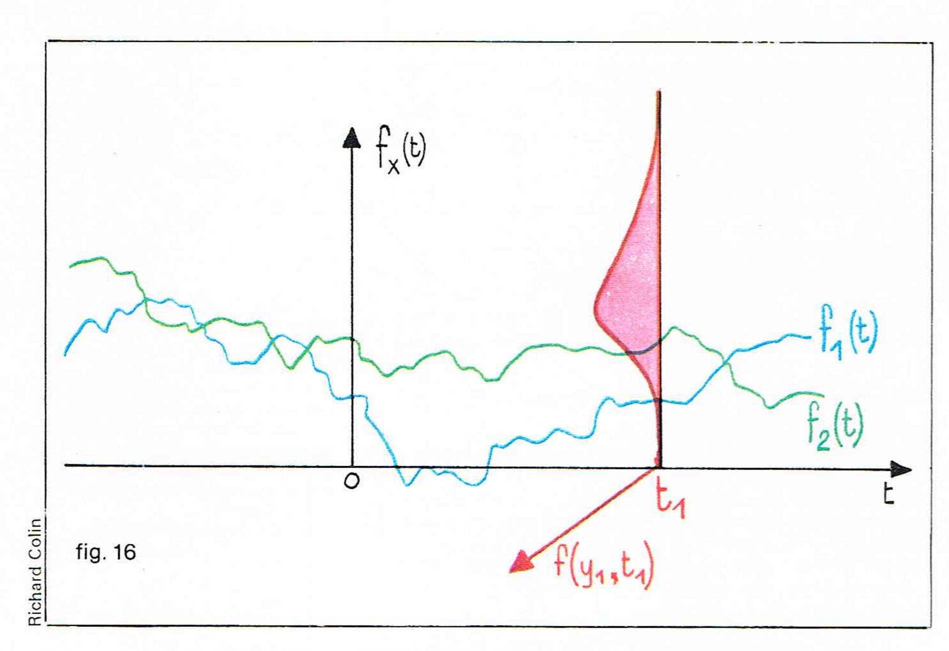
$$2 r_i = r_{i+1} + r_{i-1}$$

et, compte tenu de **(68)**, admet pour solution $r_i = \dfrac{N - i}{N - 1}$

Naturellement, la probabilité pour que Pépita rentre dans la cage située à la position N est $1 - r_i$.

Un problème tout à fait semblable est celui de deux joueurs A et B. La mise initiale de A est de i francs, celle de B est $N - i$ francs. A chaque partie, A a la probabilité p de gagner 1 franc, et q de perdre 1 franc. La probabilité pour que, après un nombre quelconque de parties, A soit ruiné est évidemment donnée par **(69)**, après substitution de $N - 1$ par N, et $i - 1$ par i, car aux N états correspondant au capital 1, ...N francs pour A il faut ajouter l'état pour lequel ce capital est de 0 franc.

Prenons pour terminer un exemple d'application des chaînes de Markov à la génétique. Supposons qu'un



▲ *Figure 16 : deux exemples de fonctions aléatoires dépendant d'un seul paramètre* t.

certain caractère soit lié aux chromosomes d'une certaine paire. Appelons D tout chromosome portant ce caractère, et *d* tout chromosome qui ne le porte pas. La paire de chromosomes en question, chez un individu donné, peut donc être l'une quelconque des combinaisons DD (état « dominant »), D*d* (état « hybride ») ou *dd* (état « récessif »).

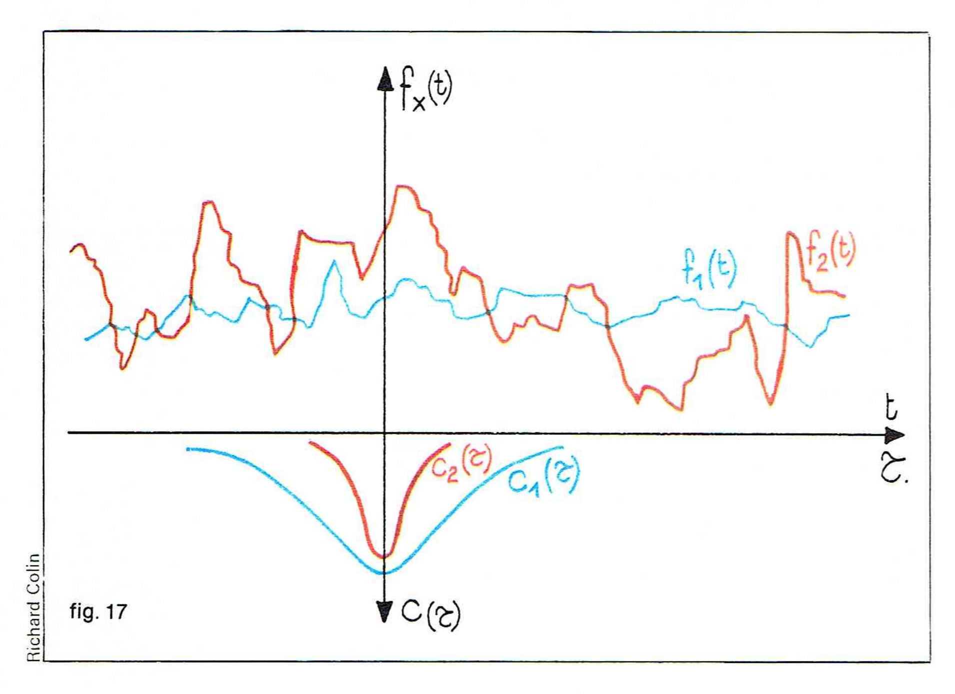
Le croisement d'un individu dominant (DD) avec un individu hybride (D*d*) peut donner lieu, comme on le voit aisément, aux combinaisons suivantes chez leurs enfants : DD $\left(\text{probabilité } \frac{1}{2}\right)$ ou D*d* $\left(\text{probabilité } \frac{1}{2}\right)$.

Celui d'un individu hybride avec un individu hybride peut donner lieu aux combinaisons suivantes : DD $\left(\frac{1}{4}\right)$, D*d* $\left(\frac{1}{2}\right)$, *dd* $\left(\frac{1}{4}\right)$.

Celui d'un individu récessif avec un individu hybride peut donner lieu aux combinaisons suivantes : D*d* $\left(\frac{1}{2}\right)$, *dd* $\left(\frac{1}{2}\right)$.

Au total, le fait de croiser un individu quelconque avec un individu hybride est un processus markovien dont la matrice stochastique est :

▼ *Figure 17 : voir développement dans le texte page 189.*



$$\underset{\sim}{P} = \begin{vmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{vmatrix}$$

Si nous croisons de nouveau les enfants de la 1re génération avec un individu hybride, les probabilités d'obtenir les états dominants, hybrides ou récessifs à la deuxième génération sont données par :

$$\underset{\sim}{P}^2 = \begin{vmatrix} \frac{3}{8} & \frac{1}{2} & \frac{1}{8} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{8} & \frac{1}{2} & \frac{3}{8} \end{vmatrix}$$

qui ne possède pas d'élément nul. Le processus markovien en question est donc une chaîne régulière. Sa limite pour un nombre très grand de générations est évidemment :

$$\lim_{n \to \infty} \underset{\sim}{P}^n = \begin{vmatrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{vmatrix}$$

car les multiplications successives de $\underset{\sim}{P}$ par lui-même préservent la 2e ligne de $\underset{\sim}{P}$. On a donc le résultat suivant : quel que soit l'état de l'aïeul que nous avons croisé avec un hybride, si toutes les générations successives sont systématiquement croisées avec des hybrides, il s'établit au bout d'un grand nombre de générations un équilibre statistique tel que $\frac{1}{4}$ des enfants sont du type dominant, $\frac{1}{4}$ du type récessif et $\frac{1}{2}$ du type hybride.

Fonctions aléatoires

Une fonction aléatoire est une fonction $f_X(t)$ dont la forme dépend du résultat du tirage d'une (ou plusieurs) variable aléatoire X. Les fonctions aléatoires peuvent être définies sur le corps des réels ou sur celui des complexes ; et elles peuvent dépendre d'un seul ou de plusieurs paramètres *t*. Nous ne nous préoccuperons ici que des fonctions aléatoires à valeurs réelles, dépendant d'un seul paramètre *t*. Le plus souvent, ce paramètre est le temps (auquel cas la fonction aléatoire est à proprement parler un processus stochastique), mais cela n'est pas nécessaire.

La *figure* 16 présente deux exemples de fonctions aléatoires de ce type. La fonction $f_1(t)$ est la fonction obtenue quand l'événement $X = x_1$ est réalisé, et la fonction $f_2(t)$ est la fonction obtenue quand l'événement $X = x_2$ est réalisé.

A un instant $t = t_1$ fixé, $Y = f_X(t_1)$ est une variable aléatoire. Appelons $f(y_1; t_1)$ la densité qui lui est associée. Les propriétés statistiques de la fonction aléatoire sont définies par la connaissance de toutes les densités de probabilité :

(70) $$f(y_1, y_2, ..., y_N;\ t_1, t_2, ..., t_N)$$

pour que Y prenne la valeur y_1 en t_1, y_2 en t_2 ... y_N en t_N pour un ensemble quelconque d'instants $t_1, ..., t_N$.

Naturellement, on peut aussi considérer la fonction caractéristique :

$$\varphi(u_1, ..., u_N; t_1, ..., t_N) = \int e^{i(u_1 y_1 + ... u_N y_N)} \times f(y_1, ..., y_N; t_1, ..., t_N)\, dy_1, ..., dy_N$$

dont la connaissance équivaut à celle de

$$f(y_1, ..., y_N; t_1, ..., t_N)$$

[voir **(26)**].

Les *moments* les plus simples associés à **(70)** sont :

— l'*espérance mathématique*

$$\langle y(t) \rangle = \int y f(y; t)\, dy ;$$

— la *covariance*

$$\langle y_1(t_1)\, y_2(t_2)\rangle = \int y_1 y_2 f(y_1, y_2; t_1, t_2)\, dy_1 dy_2$$

qui donne, en faisant $t_1 = t_2 = t$, le moment d'ordre 2 associé à $f(y; t)$.

Fonctions aléatoires stationnaires

Une fonction aléatoire $f_X(t)$ est dite stationnaire si toutes ses propriétés statistiques sont invariantes pour toute translation de l'axe des temps. Autrement dit $f_X(t)$ est stationnaire si les densités de probabilités **(70)** ne dépendent que des différences $t_2 - t_1, ..., t_N - t_1$ et non de la valeur particulière t_1 :

$$f(y_1, y_2, ..., y_N; t_1, t_2, ..., t_N) = f(y_1, y_2, ..., y_N; 0, t_2 - t_1, ..., t_N - t_1)\ \forall t_1.$$

Dans ce cas, le moment d'ordre 1, $\langle y(t)\rangle$, ne peut être qu'une constante indépendante de t. La covariance est une fonction de la différence $\tau = t_2 - t_1$. Elle s'appelle la *fonction de corrélation* de $f_X(t)$:

$$C(\tau) = \langle y_1(t)\, y_2(t - \tau)\rangle$$

Il est clair que $C(\tau)$ est une fonction paire de τ :

$$C(\tau) = \langle y_1(t)\, y_2(t - \tau)\rangle = \langle y_1(t - \tau)\, y_2(t)\rangle = C(-\tau)$$

et qu'elle est maximale pour $\tau = 0$:

$$|C(\tau)| \leqslant C(0) = \langle y^2(t)\rangle$$

En effet le coefficient de corrélation entre $y(t)$ et $y(t + \tau)$ est : d'après **(47)**

$$\rho = \frac{\langle y(t)\, y(t + \tau)\rangle - \langle y(t)\rangle^2}{\langle y^2(t)\rangle - \langle y(t)\rangle^2} = \frac{C(\tau) - \langle y(t)\rangle^2}{C(0) - \langle y(t)\rangle^2} \leqslant 1,$$

La fonction de corrélation mesure donc, en quelque sorte, la « rigidité » de la fonction aléatoire $f_X(t)$. Moins la valeur $y(t)$ prise par cette dernière à l'instant t influe sur les valeurs qu'elle peut prendre à l'instant voisin $t + \tau$, plus la fonction de corrélation tend rapidement vers 0 (*fig.* 17).

Si la fonction de corrélation $C(\tau)$ est continue à l'origine, la fonction aléatoire $f_X(t)$ est *continue en moyenne quadratique*, c'est-à-dire que l'on a :

$$\lim_{\tau \to 0} \langle [y(t + \tau) - y(t)]^2\rangle = 2\langle y^2(t)\rangle - 2\langle y(t + \tau)\, y(t)\rangle = 2[C(0) - C(\tau)] = 0.$$

Une fonction aléatoire dont l'espérance mathématique $\langle y(t)\rangle = m$ est stationnaire, et dont la covariance $\langle y_1(t_1)\, y_2(t_2)\rangle = C(\tau = t_1 - t_2)$ est continue à l'origine et est également stationnaire, est dite fonction aléatoire *stationnaire d'ordre deux*. Une telle fonction, dont les premiers moments sont stationnaires, n'est pas nécessairement stationnaire au sens strict puisque nous n'imposons pas de condition de stationnarité pour les moments d'ordre supérieur. Cependant, une *fonction aléatoire gaussienne*, c'est-à-dire telle que toutes les densités de probabilités **(70)** sont des distributions normales à N dimensions, est stationnaire au sens strict si elle est stationnaire d'ordre deux. Cela résulte du fait que tous les moments d'une distribution normale sont fixés, si les deux premiers moments le sont.

Analyse harmonique

Dans de très nombreux arrangements expérimentaux de physique (en électronique et électrotechnique en particulier) les appareils utilisés transforment la fonction aléatoire $y = f_X(t)$, dite signal d'entrée, en une autre fonction $z = \varphi_X(t)$, dite signal de sortie. La transformation $\mathcal{F} : f_X(t) \to \varphi_X(t)$ est le plus souvent linéaire et homogène, c'est-à-dire qu'elle obéit aux relations :

$$\textbf{(71)}\quad \begin{cases} \lambda f_X(t) \to \lambda \varphi_X(t) \quad \forall \lambda \text{ réel ou complexe} \\ f_X(t) + f_Y(t) \to \varphi_X(t) + \varphi_Y(t) \end{cases}$$

$$\textbf{(72)}\quad f_X(t - t_0) \to \varphi_X(t - t_0)$$

Les fonctions exponentielles $e^{i\omega t}$ sont les fonctions propres de telles transformations :

$$\textbf{(73)}\quad e^{i\omega t} \to \lambda e^{i\omega t}.$$

Pour cette raison, il est extrêmement important de savoir décomposer un signal d'entrée en somme d'exponentielles complexes (ou, de façon équivalente, en somme de fonctions sinus). C'est ce que l'on appelle faire l'analyse harmonique du signal d'entrée.

Vérifions la propriété **(73)**. Pour cela, appelons $R(t)$ la transformée de la *fonction de Dirac* $\delta(t)$ dans la transformation linéaire $\mathcal{F}$. Étant donné **(30)** :

$$y(t) = \int_{-\infty}^{+\infty} y(\theta)\, \delta(t - \theta)\, d\theta$$

nous avons, par suite de **(71)** et **(72)** :

$$\textbf{(74)}\quad z(t) = \int_{-\infty}^{+\infty} y(\theta)\, R(t - \theta)\, d\theta$$

Si $y(\theta) = e^{i\omega\theta}$, nous avons :

$$\textbf{(75)}\quad z(t) = \int_{-\infty}^{+\infty} e^{i\omega\theta} R(t - \theta)\, d\theta = g(\omega)\, e^{i\omega t}$$

où

$$\textbf{(76)}\quad g(\omega) = \int_{-\infty}^{+\infty} e^{-i\omega t'} R(t')\, dt'$$

est appelé le *gain* de la transformation $\mathcal{F}$. L'équation **(75)** est bien identique à **(73)** avec $\lambda = g(\omega)$.

La *transformation de Fourier*, déjà introduite en **(25)** est à la base de l'analyse harmonique, puisqu'elle consiste à écrire $y(t)$ sous la forme de sommes d'exponentielles $e^{i\omega t}$. On a :

$$y(t) = \int_{-\infty}^{+\infty} e^{i\omega t}\, \eta(\nu)\, d\nu$$

$$\eta(\nu) = \int_{-\infty}^{+\infty} e^{-i\omega t}\, y(t)\, dt$$

où $\nu = \dfrac{\omega}{2\pi}$. Le gain **(76)** est donc la transformée de Fourier de la « réponse percussionnelle », c'est-à-dire de la réponse à un signal en forme de fonction de Dirac.

Un *filtre passe-bande* est caractérisé par un gain tel que

$$g(\nu) = 0 \quad \text{si } \nu < \nu_1 \text{ ou } \nu > \nu_2$$
$$g(\nu) = 1 \quad \nu_1 \leqslant \nu \leqslant \nu_2$$

La réponse percussionnelle correspondante est :

$$R(t) = e^{i\frac{\omega_1 + \omega_2}{2}t}\, \frac{\sin \pi(\nu_2 - \nu_1)t}{\pi t}$$

Supposons que la fonction de corrélation $C(\tau)$ du signal d'entrée possède une transformée de Fourier $\gamma(\nu)$:

$$\textbf{(77)}\quad \gamma(\nu) = \int_{-\infty}^{+\infty} e^{-i\omega\tau}\, C(\tau)\, d\tau$$

$$\textbf{(78)}\quad C(\tau) = \int_{-\infty}^{+\infty} e^{i\omega\tau}\, \gamma(\nu)\, d\nu.$$

La transformée de Fourier de la fonction de corrélation est une fonction réelle et paire, parce que $C(\tau)$ est réelle et paire. De plus, elle est toujours positive, comme cela résulte de l'importante propriété suivante :

$$\textbf{(79)}\quad \langle |z(t)|^2\rangle = \int_{-\infty}^{+\infty} \gamma(\nu)\, |g(\nu)|^2\, d\nu.$$

En effet : $\langle |z(t)|^2\rangle$

$$= \left\langle \int\!\!\int y(\theta_1)\, R(t - \theta_1)\, y(\theta_2)\, R^*(t - \theta_2)\, d\theta_1 d\theta_2 \right\rangle$$

où nous avons utilisé **(74)** et compte tenu du fait que la réponse peut être une fonction complexe.

$$\langle |z(t)|^2\rangle = \int\!\!\int \langle y(\theta_1)\, y(\theta_2)\rangle\, R(t - \theta_1)\, R^*(t - \theta_2)\, d\theta_1\, d\theta_2$$

$$= \int\!\!\int C(\theta_1 - \theta_2)\, R(t - \theta_1)\, R^*(t - \theta_2)\, d\theta_1\, d\theta_2$$

Introduisant **(77)** :

$$\langle |z(t)|^2\rangle = \int\!\!\int\!\!\int e^{i\omega(\theta_1 - \theta_2)}\, \gamma(\nu)\, R(t - \theta_1)\, R^*(t - \theta_2)\, d\theta_1\, d\theta_2\, d\nu$$

$$= \int \gamma(\nu) \left[\int R(t - \theta_1)\, e^{-i\omega(t - \theta_1)}\, d\theta_1\right] \left[\int R^*(t - \theta_2)\, e^{i\omega(t - \theta_2)}\, d\theta_2\right] d\nu$$

$$= \int_{-\infty}^{+\infty} \gamma(\nu)\, |g(\nu)|^2\, d\nu.$$


Parimage - Caméra Press

▲ ***Le jeu de la roulette : l'erreur, ici, est de confondre les notions de*** **fréquence et de nombre de cas favorables.** ***L'écart type sur la première diminue avec le nombre de coups joués, mais celui sur le second augmente.***

Enfin, de **(78)** et **(79)**, on déduit que :

$$\textbf{(80)} \qquad \langle y^2(t) \rangle = \mathrm{C}(0) = \int_{-\infty}^{+\infty} \gamma(\nu)\, d\nu,$$

$$\textbf{(81)} \qquad \langle |z(t)|^2 \rangle = \int_{\nu_1}^{\nu_2} \gamma(\nu)\, d\nu,$$

où $z(t)$ est la réponse obtenue après passage dans un filtre passe-bande pour la bande de fréquence (ν_1, ν_2), par suite de la définition du gain correspondant. Les formules **(79)**, **(80)** et **(81)** relient les puissances moyennes des signaux d'entrée et de sortie à la connaissance de la transformée de Fourier $\gamma(\nu)$ de la fonction de corrélation.

Quelques considérations sur les probabilités

Les probabilités et la vie courante

La mémoire de la chance

Le hasard n'a ni conscience ni mémoire. Cette vérité d'allure banale est pourtant à l'origine d'un des aspects les plus paradoxaux de la théorie des probabilités, celui qui heurte le plus le sens commun.

Il est difficile à quelqu'un qui joue à la roulette d'admettre que, s'il a observé depuis le début de la partie une série de cinq coups pendant lesquels « rouge » est toujours sorti, la probabilité d'obtenir « rouge » au sixième est encore exactement égale à la probabilité d'obtenir « noir ». Il lui semble que, puisque au cours d'une très longue série de coups les nombres de « rouges » et « noirs » sont toujours à peu près égaux, si on a observé au départ une longue série de « rouges », la roulette doit se rattraper en jouant « noir » plus souvent par la suite.

L'erreur, ici, est de confondre les notions de *fréquence et de probabilité*, d'une part, et de *nombre de fois que « rouge » et « noir » sortent dans une partie*, de l'autre. Il est parfaitement exact que, conformément à la deuxième loi des grands nombres, dans une suite de tirages répétés et mutuellement indépendants, la fréquence doit tendre vers la probabilité et que l'écart type sur la *fréquence* diminue quand le nombre de coups joués augmente. Mais l'écart type sur le *nombre* de fois que l'on obtient « rouge », ou « noir », lui, augmente comme la racine carrée du nombre de coups joués. Un grand écart par rapport à l'espérance mathématique dans une première série de coups n'implique pas un écart en sens contraire dans la deuxième série. Les écarts types des diverses séries s'ajoutent toujours quadratiquement, conformément à la relation **(35)**, quand les séries sont mutuellement indépendantes.

Pourtant le type de raisonnement du joueur de roulette ci-dessus est extrêmement fréquent. Citons ainsi la réflexion selon laquelle, puisqu'on n'a pas eu d'accident de circulation avec sa voiture depuis longtemps, on serait plus particulièrement menacé d'en avoir un dans un proche avenir, ou encore celle selon laquelle, étant donné qu'il fait mauvais depuis quarante jours, la chance devient plus grande pour qu'il fasse beau le lendemain, ne serait-ce que pour vérifier l'adage « après la pluie le beau temps ». Dans ces deux derniers exemples, c'est le contraire qui est vrai, car les épreuves successives ne sont pas absolument indépendantes, il existe un coefficient de corrélation *positif* entre eux (une journée de mauvais temps, au moins sous les climats tempérés, entraîne une plus grande probabilité de mauvais temps pour le lendemain).

Les probabilités négligeables

Un deuxième aspect subjectif de la théorie des probabilités provient de la difficulté d'appréhender correctement les chances qu'un événement qui a une très petite probabilité de se produire se produise effectivement. E. Borel parle à ce sujet de la « loi unique du hasard », loi que l'on peut formuler ainsi : un événement qui a une probabilité négligeable ne se produit pas. Pourtant on entend fréquemment autour de nous des réflexions du type suivant : puisqu'il y a cent mille billets à cette tombola, j'ai une chance sur cent mille de gagner le gros lot. Comme de toutes façons *quelqu'un* gagne ce lot, pourquoi ne serait-ce pas moi ? Ainsi se persuade-t-on d'acheter un billet. Du point de vue de la seule théorie des probabilités, cependant, il faudrait jouer cent mille fois à cette tombola (ce qui est impossible au cours d'une vie) pour avoir une chance sérieuse de remporter ce gros lot.

Ainsi, à la question : à partir de quel niveau peut-on considérer une probabilité comme négligeable ? la théorie des probabilités répond : quand le nombre d'épreuves est très petit devant l'inverse de la probabilité. Le nombre d'épreuves varie énormément suivant la nature de l'événement considéré. Il est donc utile de fixer quelques ordres de grandeur.

— La vie d'un homme adulte est d'environ 20 000 journées. Nous pourrons donc, chacun pour soi, légitimement vivre sans souci des événements qui pour-

raient nous arriver, et dont la probabilité de se produire, par personne et par jour, est petite devant 1/20 000 ; par exemple : gagner le gros lot d'une loterie de cent mille billets (probabilité 10^{-5}, s'il existe un tirage par jour) ; avoir un accident mortel de la circulation à Paris (probabilité de l'ordre de 10^{-6}), etc.

— La population de la planète est de quelques milliards, et si nous cumulons toutes les générations passées, le nombre d'êtres humains ayant vécu sur terre reste inférieur à 10^{10}. Par conséquent, nous pouvons négliger les chances qu'un événement dont la probabilité de se produire, par personne et par jour, est petite devant 10^{-15}, soit arrivé une fois dans l'histoire humaine. Telle est, par exemple, la probabilité d'obtenir 50 fois « pile » en suivant au jeu de pile ou face. Considérant que tous les hommes de tous les temps n'ont certainement pas joué en moyenne une partie de cinquante coups de pile ou face chaque jour, il est pratiquement exclu qu'un tel événement se soit effectivement produit, même une seule fois.

— Certaines lois de la physique, en particulier les lois de la thermodynamique, sont des lois statistiques. Le deuxième principe de la thermodynamique, par exemple, nous enseigne qu'une fois l'équilibre atteint, un litre de gaz occupant un ballon d'un litre l'occupe uniformément, les molécules ne se groupant jamais, par exemple, dans la moitié inférieure du ballon et laissant l'autre moitié vide. Évaluons la probabilité de mettre en évidence, par exemple, une fluctuation de densité ou de pression entre la moitié supérieure et la moitié inférieure du ballon, à la précision relative extraordinaire de 10^{-9} (un milliardième). Il y a dans le ballon environ $4 \cdot 10^{22}$ molécules de gaz, chacune d'elles ayant la probabilité $\frac{1}{2}$ de se trouver dans l'une ou l'autre moitié du récipient. L'équation **(38)** nous apprend que la distribution de probabilité d'avoir N_A molécules dans la moitié inférieure est

$$N\ (N_A - 2 \cdot 10^{22}, 10^{11}).$$

Une fluctuation relative supérieure à 10^{-9} correspond à $N_A < 2 \cdot 10^{22} - 2 \cdot 10^{13}$ ou $N_A > 2 \cdot 10^{22} + 2 \cdot 10^{13}$, c'est-à-dire à un écart supérieur à 200 écarts types ($\sigma = 10^{11}$). La probabilité correspondante est beaucoup plus petite que 10^{-15}.

— On estime généralement que le nombre d'atomes de matière dans l'univers est inférieur à 10^{100} atomes, et qu'il s'est écoulé, depuis le big bang primitif, environ 12 milliards d'années, soit un temps inférieur à 10^{18} s. Tout phénomène physique de l'échelle atomique qui aurait une probabilité de se produire inférieure ou de l'ordre de 10^{-120} par seconde n'a donc pratiquement aucune chance de s'être jamais réalisé dans l'univers.

Probabilités d'ignorance et indéterminisme fondamental

Dans la plupart des applications de la théorie des probabilités, l'événement aléatoire survient au terme d'une succession de causes physiques qui entraîne inéluctablement son apparition. Ainsi, connaissant parfaitement la position du dé au moment où il est lancé, l'impulsion que lui donne le joueur, et la géométrie et les conditions de rugosité des faces du dé et de la table, il est, en principe, possible de prédire quel sera le numéro tiré. Mais l'écheveau des causes est si embrouillé qu'il est impossible de faire le calcul. Dans ce cas, l'utilisation du calcul des probabilités ne fait donc que cacher notre incapacité pratique de calculer la trajectoire du dé. Bien entendu, le plus souvent, certaines données des problèmes posés nous échappent. Ainsi, nous ne connaissons pas exactement la position et l'impulsion initiales du dé, etc. Dans les problèmes de physique statistique, notre ignorance est encore plus grande : nous ignorons généralement tout de la position et de l'impulsion initiales de chacune des $4 \cdot 10^{22}$ molécules d'un litre de gaz. Dans les jeux de hasard, cette ignorance est parfois volontaire (les cartes sont distribuées retournées). Bref, dans tous ces cas, on parle de *probabilité d'ignorance*.

Notre ignorance des paramètres du problème n'exclut nullement que les grandeurs physiques correspondantes aient réellement une valeur, mais nous ne voulons pas ou nous ne pouvons pas les mesurer. L'efficacité du calcul des probabilités, dans ce cas, tire souvent avantage de la multiplicité même des causes pouvant influer sur le résultat final. Cette multiplicité détermine généralement les conditions d'applications des lois des grands nombres.

Mais l'on sait, depuis 1927, que l'ignorance n'est pas le seul fondement possible du concept de probabilité. L'avènement de la mécanique quantique a montré, au contraire, qu'il est impossible d'assigner une valeur précise à la plupart des grandeurs physiques généralement attachées à un système physique concret (atomes, molécules, etc.). En particulier, il est impossible de construire un arrangement expérimental tel qu'à la fois la *position* et la *vitesse* d'un atome aient une signification précise. Le résultat de la mesure de telles grandeurs est le résultat d'un tirage au hasard, qui n'est déterminé par aucune cause sous-jacente (pour autant, du moins, que la théorie physique actuelle soit correcte). Ce hasard est le reflet d'une solution de continuité fondamentale dans les processus de mesure. On dit, non sans un certain excès de langage, qu'il s'agit d'un *indéterminisme fondamental*. Comme tous les corps, même ceux à notre échelle, même ceux à l'échelle astronomique, obéissent, en dernier ressort, aux lois de la mécanique quantique, on voit que le concept de probabilité a sans doute des racines plus profondes et une portée beaucoup plus large que la simple probabilité d'ignorance, par laquelle il a été historiquement fondé.

Quelques exemples

Le paradoxe de Saint-Pétersbourg

Pour se libérer d'une grosse dette envers Paul, Pierre envisage de jouer avec lui au jeu de pile ou face, de la façon suivante : au premier coup, si la pièce montre « pile » le jeu s'arrête et Pierre est libéré de sa dette, mais si « face » sort, Pierre donne deux francs à Paul et le jeu continue une autre fois ; si, au deuxième coup, « pile » sort, le jeu s'arrête ; si « face » sort, Pierre donne quatre francs à Paul et le jeu continue encore.

A chaque partie perdue par Pierre, celui-ci donne à Paul le double de la somme qu'il lui a donnée à la partie précédente. Le jeu s'arrête dès que, enfin, « pile » sort.

Afin de connaître si cette façon de jouer est avantageuse pour lui, Pierre calcule l'espérance de gain de Paul. Il trouve que cette espérance est de un franc pour la première partie, de un franc encore pour la seconde (une chance sur deux de jouer cette partie et une chance sur deux de gagner 4 francs), de un franc encore pour la troisième, et ainsi de suite... Au total, l'espérance de gain de Paul est infinie. Aucune dette ne vaut la peine d'accepter les règles de ce jeu.

En pratique cependant, la réalité est fort différente. Si Pierre perd les *n* premières parties, il doit donner à Paul, partie après partie, une somme totale de

$$2 + 4 + \ldots 2^n = 2^{n+1} - 2 \text{ francs.}$$

Il ne peut se permettre de perdre qu'un petit nombre de fois, faute de quoi il se trouve bientôt insolvable. Possédât-il 10 millions de francs, il ne pourrait perdre sans faire faillite que 22 fois en suivant. L'espérance de gain de Paul, en définitive, n'est donc que de 22 francs.

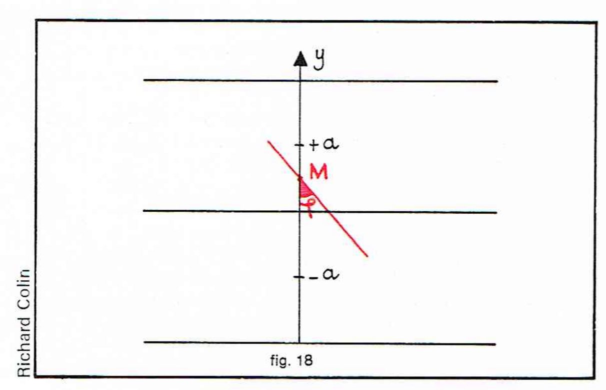
L'exemple ci-dessus nous amène à méditer sur le comportement des personnes qui, jouant aux jeux de hasard (roulette), appliquent la tactique suivante : en cas de perte, doubler la mise avec obstination et rejouer, jusqu'au succès final. En réalité, même si, dans un jeu parfaitement équitable, le joueur qui applique cette tactique est assuré de finir par équilibrer ses pertes, voire d'obtenir un bénéfice, l'avantage décisif de la banque vient de ce qu'elle est normalement capable de pousser le jeu plus loin du côté des pertes que les joueurs. Ceux-ci sont acculés à la faillite avant elle, et, par suite, l'espérance de gain de la banque est en faveur de cette dernière.

Le problème de l'aiguille (Buffon)

Supposons que nous ayons un parquet, fait de lames de bois parallèles de largeur $2a$, et que nous lancions sur ce parquet, au hasard, une aiguille rectiligne de longueur $2l < 2a$. Quelle est la probabilité pour que l'aiguille coupe une des raies du parquet ?

On voit immédiatement sur la *figure* 18 que le problème dépend de deux variables aléatoires indépendantes : l'ordonnée y ($-a \leqslant y < a$) du point M, milieu de

▶ *Figure 18 : le problème de l'aiguille (Buffon).*



l'aiguille, par rapport à la ligne de parquet la plus proche, et l'angle φ $\left(-\frac{\pi}{2} \leqslant \varphi < \frac{\pi}{2}\right)$ que fait l'aiguille avec la perpendiculaire tracée de M sur les lignes du parquet. Ces deux variables aléatoires admettent des densités de probabilités uniformes :

$$f_1(y) = \frac{1}{2a} dy$$

$$f_2(\varphi) = \frac{1}{\pi} d\varphi$$

comme $l < a$, l'aiguille ne peut couper que la raie considérée. La condition nécessaire et suffisante pour cela est $|y| \leqslant l \cos \varphi$.

Par **(3)**, la probabilité p cherchée est

$$p = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} f_2(\varphi)\, d\varphi \int_{-l\cos\varphi}^{l\cos\varphi} f_1(y)\, dy = \frac{2l}{a\pi}$$

En son temps (le XVII[e] siècle), ce résultat a beaucoup intrigué les philosophes. Comment un problème dans lequel il n'est question que de réseau et de segments rectilignes peut-il engendrer le nombre π ? Quel rapport ce nombre entretient-il avec les fondements de la théorie des probabilités ? Cette curiosité, toutefois, est sans objet. Le nombre π, comme le prouve par exemple le développement en série de Taylor de la fonction $y = \text{arc tg } x$, se retrouve comme valeur de beaucoup de séries de fractions rationnelles. Le problème de l'aiguille reste néanmoins une curieuse et amusante façon de mesurer ce nombre. L'astronome Wolf, en 1850, fit 5 000 épreuves et mesura la fréquence des succès. Il en déduisit $\pi = 3{,}1596$. L'écart observé est tout à fait compatible avec la distribution des résultats de mesure prévue par la deuxième loi des grands nombres.

Des garçons et des filles

Soit le problème suivant (Gamow) : dans un pays où la polygamie est admise, un édit indique que les femmes ayant accouché d'un garçon doivent être stérilisées ; limitant ainsi le nombre de garçons, la loi espère augmenter la proportion de filles dans la nation. Dans ce pays, il n'y a donc que des familles avec un garçon, une fille et un garçon, deux filles et un garçon, ... n filles et un garçon. La pratique de la stérilisation limite évidemment le nombre moyen d'enfants de chaque famille. Augmente-t-elle cependant, comme on pourrait le croire, la proportion de filles ?

La réponse est non. Car, au premier accouchement, toutes les femmes ont autant de chances de mettre au monde un garçon qu'une fille. Et à l'accouchement suivant (qui ne concerne plus que la moitié des femmes, celles qui ont une fille première-née), les mêmes ont encore une probabilité $\frac{1}{2}$ d'avoir un garçon ou une fille. Et ainsi de suite.

Le nombre moyen de garçons par famille avec enfants est évidemment un. Celui des filles, malgré les apparences, est un également : les familles avec plusieurs filles ne font que compenser les familles à fils unique. On a :

$$\langle n_F \rangle = \sum_{n=2}^{\infty} \frac{n-1}{2^n} = 1.$$

Les probabilités en biologie

Depuis les travaux de Mendel (1865), les biologistes sont convaincus de l'importance de la théorie des probabilités pour l'explication d'un très grand nombre de phénomènes biologiques tels que la reproduction, les mutations génétiques dans les chromosomes, la réplication de ceux-ci dans la division cellulaire, etc. L'intervention du hasard dans les phénomènes vitaux est même si grande que certains voient en lui l'explication fondamentale des phénomènes de la vie et de l'évolution (par exemple, le livre de Jacques Monod, *le Hasard et la Nécessité*).

La génétique et ses deux grandes branches, la génétique des populations d'une part, la génétique évolutive de l'autre, font donc un très large appel aux techniques statistiques.

A partir d'un petit nombre de constituants (quatre atomes essentiels, carbone, oxygène, hydrogène, azote), on fabrique par combinaison un nombre plus grand de briques fondamentales de la matière vivante : les 20 acides aminés. Quatre molécules, toujours les mêmes, constituent les barreaux des longues chaînes de la double hélice de l'acide désoxyribonucléique, qui peut comporter jusqu'à 10^7 ou 10^8 échelons. L'ADN présent dans les chromosomes détermine le patrimoine héréditaire. Celui-ci est préservé dans la réplication, phénomène intervenant dans la division cellulaire, et combiné dans la reproduction. Dans ce dernier phénomène, le patrimoine héréditaire d'un individu est combiné (généralement à parts égales, mais parfois non, comme dans la reproduction bactérienne) avec le patrimoine héréditaire d'un autre individu, sur la base du hasard.

Le hasard intervient aussi dans les phénomènes de mutation, c'est-à-dire de changement dans l'ordre ou dans la structure des barreaux de l'ADN. Ces mutations peuvent être induites (elles peuvent par exemple être le résultat accidentel du passage d'un rayon cosmique) ou spontanées, en particulier au cours de la réplication. Certains pensent même que les mutations spontanées pourraient avoir une relation avec l'indéterminisme fondamental dont nous avons parlé plus haut, la fonction d'onde de la molécule d'ADN pouvant assumer virtuellement plusieurs états avec des probabilités de réalisation pour chacun d'eux. Quoi qu'il en soit, il existe donc en général une probabilité petite, mais non nulle, de transformation génétique à chaque étape de la division cellulaire. La stabilité du patrimoine héréditaire ne peut donc s'expliquer que par l'intervention d'équilibres entre des transitions en sens inverses, comme nous en avons mentionné dans l'étude des processus markoviens.

Chez l'homme, le patrimoine génétique est contenu dans les 46 chromosomes. Au cours de la reproduction, 23 chromosomes du gamète mâle sont combinés deux à deux avec les 23 chromosomes du gamète femelle. Il y a donc, au total, 4^{23} ensembles complets de chromosomes, c'est-à-dire plus de 600 milliards d'enfants génétiquement différents les uns des autres, pouvant être engendrés par un même couple de parents. Seuls de vrais jumeaux, issus d'une même cellule originale, possèdent exactement les mêmes caractères génétiques. Dans tous les autres cas, il y a seulement une certaine chance, facile à calculer, pour que deux enfants, ou deux individus de la n-ième génération, aient dans leur carte génétique le même chromosome, et, partant, les mêmes caractères liés à ce chromosome.

Les données statistiques

La visualisation des données statistiques

Les réponses à une question donnée, faisant partie d'une enquête statistique, varient d'un individu à l'autre. On observe que le pourcentage de réponses « oui » fluctue selon l'échantillon de personnes interrogées. Pour avoir une idée assez précise de la *probabilité intrinsèque* (en admettant qu'elle existe) attachée à cette réponse, il faut interroger un échantillon suffisamment large : si les conditions d'application de la deuxième loi des grands nombres sont remplies, plus le nombre de personnes interrogées est important, meilleure sera l'évaluation de cette probabilité.

Le résultat de la mesure d'une grandeur physique fluctue selon le type d'appareil utilisé et l'habileté de l'expérimentateur. Une mesure unique ne donnera, le plus souvent, qu'une indication grossière sur la valeur réelle de cette grandeur physique : pour avoir une idée plus précise sur cette grandeur, il est souvent utile de procéder à une série de mesures indépendantes. Les résultats se répartissent alors de part et d'autre d'une valeur intermédiaire « moyenne ». Cette valeur est adoptée comme estimation de la grandeur, et les écarts observés autour d'elle donnent une indication sur la précision de cette estimation : plus les mesures sont dispersées, plus large est la zone d'incertitude sur la valeur réelle de la grandeur. Ici encore, plus nombreuses sont les mesures, meilleure est, en général, la précision de l'estimation.

En outre, comme on l'a vu, il existe une classe de phénomènes physiques (les phénomènes quantiques) pour lesquels la grandeur physique mesurée n'a pas de valeur définie au moment de la mesure. Ce sont seulement les probabilités respectives pour que la mesure de cette grandeur donne telle ou telle valeur qui sont définies. Pour évaluer avec précision ces probabilités, il faut donc procéder à une série nombreuse d'expériences. Chacune d'elles, de plus, est entachée d'erreurs expérimentales et la détermination des résultats de mesures possibles nécessite donc, dans ce cas, plusieurs séries de mesures.

En définitive, dans de très nombreuses situations, on a donc une statistique de résultats. Cette statistique, sous forme brute, se présente comme une suite ennuyeuse de nombres : la première mesure a donné tel résultat, la seconde tel autre, etc. Le problème est de disposer ces résultats (par exemple sous forme de graphiques) de telle sorte que l'information contenue (par exemple, la liste des résultats de mesures possibles pour une grandeur quantique et la précision de l'estimation de chacune d'elles) soit immédiatement perceptible. C'est le problème de la *visualité des données* dont nous parlerons dans cette section.

Formulons le problème particulier d'une suite de mesures d'une grandeur physique $\mathcal{M}$ (ces mesures pourront être faites chacune dans les mêmes conditions sur le même système, mais nous ne l'exigeons pas).

L'ensemble des mesures fournit une statistique de résultats. Si la première mesure conduit au résultat m_1, la seconde m_2, etc., la statistique se représente comme la séquence des nombres $(m_1, ..., m_n)$.

Il est clair, cependant, que cette statistique *n'est pas* la séquence des valeurs $(\mu_1, ..., \mu_n)$ prises par la grandeur au long des mesures, à cause des erreurs expérimentales commises. Chaque mesure est entachée d'une certaine erreur, qui, en général, varie d'une mesure à l'autre ; la séquence des μ est liée à la séquence des m par :

$$m_i = \mu_i + \delta\mu_i$$

où $\delta\mu_i$ représente l'erreur commise dans la i-ième mesure. Le problème du physicien est de déduire, dans la mesure du possible, les μ_i à partir des m_i.

Les erreurs peuvent être de deux types : *systématiques* et *statistiques*. Les erreurs systématiques découlent de certains défauts de l'appareil de mesure. Il appartient aux physiciens de chercher à les corriger en améliorant l'appareillage et en l'étalonnant avec une meilleure précision. Nous ne considérerons pas ici ces erreurs systématiques, mais le résidu irréductible des erreurs statistiques, découlant de la précision nécessairement limitée des appareils ou d'autres raisons ; ces erreurs provoquent une dispersion des résultats de mesure, chacun d'eux se présentant comme le résultat du tirage d'une variable aléatoire « mesure de $\mathcal{M}$ », et ceci même dans le cas de mesures faites dans les mêmes conditions sur le même système.

Distinguons ici trois cas. Le premier, et le plus restrictif, est celui des résultats de mesures d'une grandeur telle que :

a - $\mathcal{M}$ prend une valeur unique tout au long de la séquence de mesures : $\mu_1 = \mu_2 = ... \mu_n = \mu_0$;

b - les causes d'erreurs sont multiples et indépendantes. L'erreur totale commise est une variable aléatoire, somme d'un nombre très grand de variables aléatoires (chaque erreur particulière), chacune d'elles étant régie par une « bonne » distribution de probabilité, au sens de la seconde loi des grands nombres.

Nous appellerons C_1 la classe de statistiques qui satisfont aux conditions ci-dessus. La classe C_1 comprend les séquences de résultats de mesure d'une grandeur physique $\mathcal{M}$ effectuée dans les mêmes conditions sur le même système, pourvu que les causes d'erreurs soient multiples et indépendantes.

Ces conditions sont approximativement réalisées dans l'exemple de la mesure de la longueur d'une table à l'aide d'une règle ; parmi les causes d'erreurs, citons : la position de la règle qui n'est jamais parfaitement d'équerre avec le bord de la table, lequel n'est en outre pas parfaitement en coïncidence avec la première graduation de la règle ; l'erreur de parallaxe entre le plan de la table, celui des graduations de la règle et les yeux de l'observateur ; l'épaisseur des graduations et les erreurs d'interpolation entre, etc. Un meilleur exemple, célèbre car il a été utilisé par Bessel pour vérifier les conclusions de la théorie probabilistique des mesures, est celui de la mesure de l'ascension droite et de la déclinaison d'une étoile ; dans ce cas, en plus des causes d'erreurs du type précédent, on trouve les fluctuations de l'indice de réfraction de l'air atmosphérique, etc.

La clause **b** implique que l'erreur totale est une variable aléatoire qui suit (de très près) une distribution gaussienne ; si l'appareil est convenablement étalonné, cette distribution est centrée sur μ_0. Les mesures faites dans les mêmes conditions sur le même système sont donc caractérisées par la distribution des résultats :

(82) $$f(m) = \mathrm{N}(m - \mu_0, \sigma_0)$$

Un cas moins restrictif est représenté par la classe C_2 de statistiques de mesures d'une grandeur $\mathcal{M}$ telle que :

a - la grandeur $\mathcal{M}$ ne prend pas la même valeur au long de la séquence de mesures, mais les valeurs successives de $\mathcal{M}$ sont liées entre elles par une relation analytique connue ;

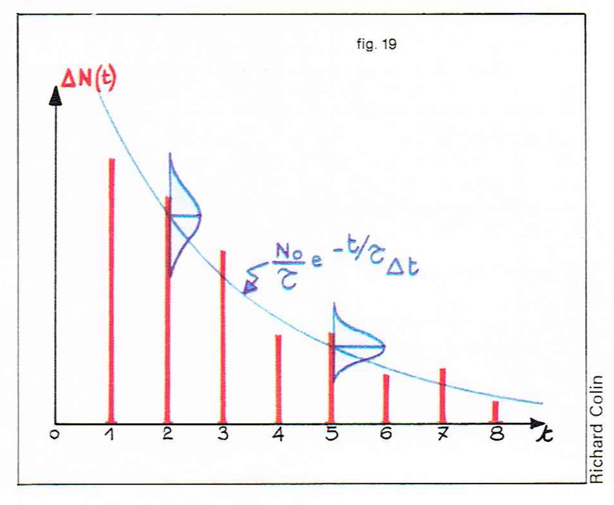
b - les mesures de $\mathcal{M}$ sont régies par une distribution gaussienne (ou approximativement gaussienne) telle que la variance de cette distribution dépend de la valeur μ prise par $\mathcal{M}$, selon une relation analytique connue, $\sigma = \sigma(\mu)$.

Un exemple de statistique de cette classe est donné par les taux de comptage dans l'observation d'un échantillon de substance radio-active. La grandeur $\mathcal{M}$ est le nombre moyen de désintégrations par unité de temps. Au cours du temps ce nombre diminue suivant une loi exponentielle, du type :

$$\Delta\overline{\mathrm{N}}(t) \simeq \frac{\mathrm{N}_0}{\tau} e^{-t/\tau} \Delta t$$

dans laquelle, en général, les paramètres τ et N_0 ne sont pas connus. Le nombre réel de désintégrations observées, $\Delta\mathrm{N}$, n'est pas le nombre $\Delta\overline{\mathrm{N}}$, à cause des fluctuations quantiques. Chaque atome de l'échantillon a une probabilité définie de se désintégrer dans l'intervalle de temps considéré, et le nombre total de désintégrations observées suit une distribution binominale — c'est-à-dire, pratiquement, une distribution de Poisson, ou encore une distribution gaussienne pourvu que $\Delta\overline{\mathrm{N}}(t) > 10$, avec $\sigma = \sqrt{Npq} \simeq \sqrt{\Delta\overline{\mathrm{N}}}(t)$ (*fig.* 19).

Notez que, dans cette expression, $\Delta\overline{\mathrm{N}}(t)$ n'est en général pas connu. Dans le cas d'une mesure isolée,



◄ ***Figure 19 : distribution binominale que suit le nombre total de désintégrations d'un échantillon de substance radio-active.***



Parimage - Caméra Press

▲ *Un exemple de visualisation des données statistiques.*

$\Delta\bar{N}$ (t) est remplacée par son évaluation ΔN (t), conformément au principe de Bayes que nous étudierons ci-après.

Finalement, considérons le cas encore plus général des statistiques de mesures où non seulement la valeur prise par la grandeur $\mathcal{M}$ mais encore les précisions de mesure varient d'une mesure à l'autre, sans qu'aucune loi de variation pour les μ ou les $\delta\mu$ ne soit connue à l'avance. On supposera cependant que la précision de l'appareil de mesure utilisé pour la i-ième mesure, soit σ_i, est connue ; cela veut dire que, considérant la i-ième mesure isolément, dont le résultat m_i a été obtenu avec un appareil ayant la précision σ_i, la probabilité ***a priori*** pour que la valeur réellement prise par $\mathcal{M}$ soit μ_i est, conformément au principe de Bayes :

(83) $$f(\mu_i) = N(m_i - \mu_i, \sigma_i)$$

[comparez avec **(82)**].

Toute l'information disponible dans une statistique de mesures de cette classe (que nous appellerons classe C_3) est l'ensemble des paires (m_i, σ_i) relatives à chaque mesure et à la précision avec laquelle cette mesure a été effectuée. Les (m_i, σ_i) peuvent être considérés comme le résultat d'un tirage effectué sur deux variables aléatoires M et Σ, généralement corrélées, et régies par une distribution de probabilité à deux dimensions ρ (m, σ).

La donnée des (m_i, σ_i) constituant la statistique des résultats de mesure suffit, dans le cas de mesures gaussiennes, à donner la totalité de l'information disponible quant à la séquence de mesures de M. Toutefois, présentée sous la forme banale d'une double liste de résultats de mesures et d'écarts types associés, cette information ne fait pas ressortir les traits essentiels de son contenu physique, et elle est peu commode, pour autant que le nombre de mesures effectuées soit assez important.

C'est la raison pour laquelle on préfère, en général, présenter de façon plus condensée et plus suggestive les résultats de mesures, fût-ce au prix d'une perte plus ou moins importante d'information.

Le diagramme (M, Σ)

La présentation la plus naturelle de cette information est le diagramme à deux dimensions (M, Σ) (*fig.* 20) ; dans ce diagramme la i-ième mesure apparaît comme un point de coordonnées (m_i, σ_i). La densité de points dans ce diagramme, conformément à la première loi des grands nombres, tend vers la densité de probabilité ρ (m, σ). Toutefois, cette représentation, bien qu'elle occasionne une perte minimale d'information (due uniquement à l'imprécision du dessin, etc.) reste peu suggestive.

Liaison entre ρ (m, σ) et φ (μ)

En fait la distribution des résultats de mesures ρ (m, σ) est liée à la distribution φ (μ) par :

(84) $$\rho(m, \sigma) = \int \varphi(\mu)\, \lambda(\sigma/\mu)\, N(m - \mu, \sigma)\, d\mu$$

car la probabilité d'obtenir pour résultat de mesure (m, σ) est, par suite de **(3)** et **(4)**, égale à la somme des probabilités pour que $\mathcal{M}$ prenne en réalité la valeur μ multipliée par la probabilité λ (σ/μ) de faire la mesure avec la précision σ, sachant que $\mathcal{M}$ a pris la valeur μ, et multipliée par la probabilité d'obtenir le résultat m, sachant que $\mathcal{M}$ a pris la valeur μ et que la mesure est faite avec la précision σ, cette dernière étant N ($m - \mu$, σ) d'après **(82)**.

Histogrammes

Dans la limite où σ est négligeable,

$$\rho(m, \sigma) \simeq \int \varphi(\mu)\, \lambda(\sigma/\mu)\, \delta(m - \mu)\, d\mu$$
$$= \varphi(m)\, \lambda(\sigma/m)$$

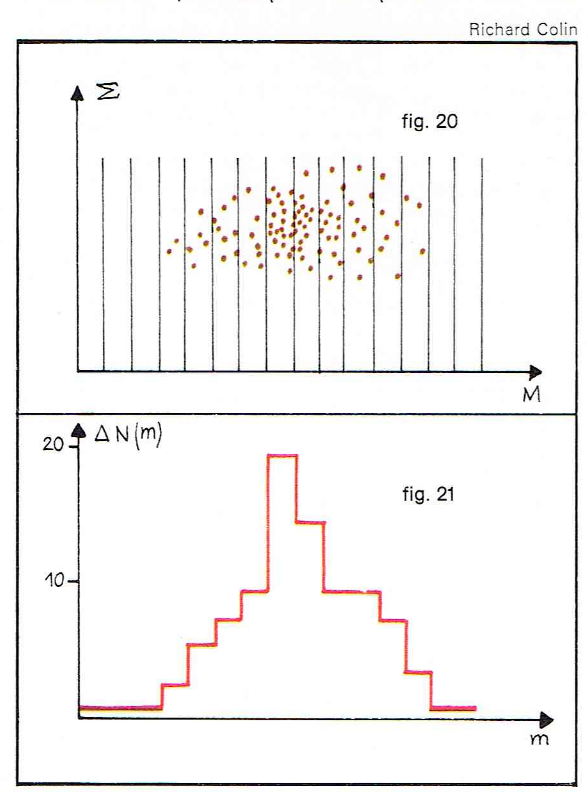
par **(30)**. La distribution φ (μ) est simplement dans ce cas la distribution marginale de ρ (m, σ). Cela nous conduit à une méthode simple, qui donne une bonne idée du spectre φ (μ) et qui est utile même dans le cas où les erreurs sont plus importantes ; cette méthode consiste à diviser l'axe M du diagramme précédent en segments, ordinairement de même longueur, et à construire ce que l'on appelle l'histogramme ΔN (m) en portant le nombre de points contenus dans chaque intervalle en fonction de m (*fig.* 21).

La présentation des résultats sous la forme d'un histogramme, quoique très simple et assez explicite, a deux inconvénients liés à deux sources de perte d'information :

— nous savons quel est le nombre total de mesures qui donne un résultat dans chacun des intervalles Δm, mais nous perdons la position exacte de chaque mesure à l'intérieur de ces intervalles ;

— chaque mesure est traitée de la même manière, quelle que soit sa précision ; l'information de cette précision est perdue.

Richard Colin



▶ *Figure 20 : le* diagramme *à deux dimensions* (M, Σ) *; figure 21 : l'histogramme* ΔN *(m).*

Afin de minimiser cette perte d'information, il convient de choisir pour taille des intervalles une longueur de l'ordre de $\langle \sigma \rangle$, l'erreur moyenne commise dans les mesures. De cette façon, il y a une bonne chance pour que les μ_i correspondant aux mesures m_i appartenant au i-ième segment appartiennent aussi au même segment ou aux segments immédiatement voisins. En outre, avec $\Delta m \simeq \langle \sigma \rangle$, il est fort probable que les mesures se distribuent de façon approximativement homogène au long de chaque segment. En effet, dans le cas extrême où la grandeur M prendrait chaque fois la même valeur μ, on observerait une dispersion des résultats de mesure due aux erreurs de mesure, telle que les densités de points aux abscisses $m = \mu + \langle \sigma \rangle$ et $m = \mu$ sont, approximativement, dans le rapport $e^{-\frac{1}{2}} \simeq 0{,}61$. La technique de l'histogramme, qui équivaut à supposer une répartition homogène des résultats de mesure le long de chaque segment, ne perd donc, en général, que peu d'information. Le choix de longueur de segment (le « bin » des Anglo-Saxons) indiqué plus haut est en définitive un bon compromis entre le désir de représenter de façon aussi immédiate que possible les données expérimentales et leur signification, et le désir de ne pas perdre l'information disponible. Néanmoins, le choix de l'intervalle peut être influencé par d'autres facteurs; notamment, quand la statistique de mesures est pauvre (les mesures peu nombreuses), il peut être préférable de prendre un intervalle de longueur Δm plus grand, soit pour mieux visualiser le spectre obtenu, soit pour se rapprocher dans chaque intervalle des conditions d'applications des lois des grands nombres.

Idéogrammes

La méthode des idéogrammes a l'avantage d'éviter la perte d'information inhérente à la méthode de l'histogramme, tout en restant très suggestive. Malheureusement, cette méthode conduit en général à des calculs peu pratiques.

L'idéogramme gaussien, le plus utilisé, est par définition la fonction aléatoire obtenue en remplaçant chaque point (m_i, σ_i) du diagramme (M, Σ) par une gaussienne $N(m - m_i, \sigma_i)$ et en faisant la somme de toutes les gaussiennes ainsi obtenues (*fig.* 22).

$$\textbf{(85)} \qquad I(m) = \sum_i N(m - m_i, \sigma_i)$$

Ainsi tient-on compte de l'abscisse exacte de chaque mesure m_i, et également de la précision associée. On doit noter, cependant, que le choix de la fonction $N(m - m_i, \sigma_i)$ dans cette technique est assez arbitraire. Dans le cas d'une mesure isolée, cette distribution représente, d'après **(83)**, la probabilité *a priori* pour que la valeur réelle de la grandeur $\mathcal{M}$ soit m, mais la somme **(85)** n'a pas d'interprétation aussi simple. En définitive, sauf dans des cas très particuliers, l'utilisation des idéogrammes gaussiens ne présente pas d'avantages décisifs sur la technique, plus simple, des histogrammes.

Problèmes d'estimation

Dans le cas plus simple de statistiques de mesures gaussiennes répétées (c'est-à-dire faites dans les mêmes conditions sur le même système), les mesures se répartissent conformément à la distribution

$$f(m) = N(m - \mu_0, \sigma_0)$$

obtenue en remplaçant, dans **(84)**,

$$\varphi(\mu) \text{ par } \delta(\mu - \mu_0) \text{ et } \lambda(\sigma/\mu) \text{ par } \delta(\sigma - \sigma_0).$$

L'estimation de $\varphi(\mu)$ se réduit, dans ce cas, à l'estimation du paramètre μ_0. Une estimation correcte de μ_0 est la moyenne arithmétique μ^* :

$$\mu^* = \frac{1}{N} \sum_i m_i$$

En effet, μ^* est une variable aléatoire dont la limite, pour $n \to \infty$, est μ_0, conformément à la première loi des grands nombres. Par ailleurs, la variance sur μ^* est, d'après **(35)** :

$$\Delta^2 = \frac{\sigma_0^2}{N}$$

Cette variance coïncide avec la valeur déduite de la seconde loi des grands nombres : c'est la plus petite possible. On dit que Δ est la précision de l'estimation.

En général σ_0 n'est pas connu, et pour évaluer Δ, on doit faire d'abord une estimation de σ_0. La moyenne arithmétique des $(m_i - \mu_0)^2$ est, selon les lois des grands nombres, une estimation correcte de σ_0.

Substituant dans cette expression μ_0, qui n'est pas connu, par μ^*, on obtient une nouvelle variable aléatoire T dont l'espérance mathématique est :

$$\begin{aligned}\langle T \rangle &= \frac{1}{N} \langle \sum_i (m_i - \mu^*)^2 \rangle \\ &= \frac{1}{N} \{N \langle m_i^2 \rangle + N \langle \mu^{*2} \rangle - 2N \langle \mu^{*2} \rangle\} \\ &= \langle m_i^2 \rangle - \langle \mu^{*2} \rangle = \sigma_0^2 + \mu_0^2 - \frac{\sigma_0^2}{N} - \mu_0^2 \\ &= \frac{N-1}{N} \sigma_0^2\end{aligned}$$

où nous avons utilisé **(34)** et **(24)**.

Choisissons alors pour estimation de σ_0 et Δ :

$$\textbf{(86)} \qquad \sigma_0^{*2} = \frac{N}{N-1} T = \frac{1}{N-1} \sum_i (m_i - \mu^*)^2$$

$$\Delta^* = \sqrt{\sum_i \frac{(m_i - \mu^*)^2}{N(N-1)}}$$

Considérons le cas plus général de statistique de mesures pour lesquelles $\varphi(\mu)$ prend encore la forme $\delta(\mu - \mu_0)$, mais $\lambda(\sigma/\mu) \neq \delta(\sigma - \sigma_0)$ (autrement dit, la grandeur $\mathcal{M}$ prend encore la valeur μ à chaque mesure, mais la précision dépend des conditions, variables, de la mesure). La formule **(84)** s'écrit dans ce cas :

$$\rho(m, \sigma) = \lambda(\sigma) \, N(m - \mu_0, \sigma),$$

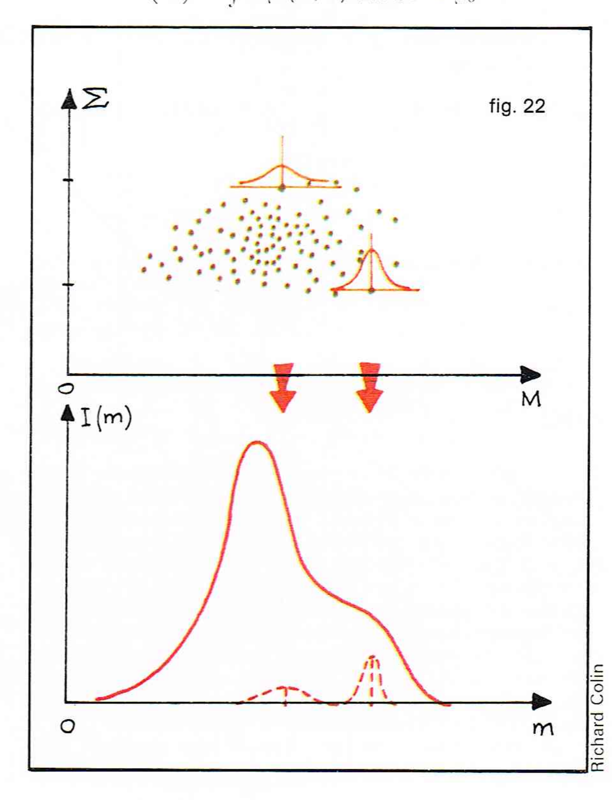
où nous écrivons $\lambda(\sigma)$ au lieu de $\lambda(\sigma/\mu_0)$, car μ_0 a une probabilité 1 de se produire.

Le problème de l'estimation du paramètre μ_0 est dans ce cas plus compliqué.

De nouvelles techniques nous aideront à trouver la solution.

On peut noter, cependant, que l'espérance mathématique de M est μ_0 :

$$\langle M \rangle = \int m \rho(m, \sigma) \, dm \, d\sigma = \mu_0$$



◄ ***Figure 22 :* l'idéogramme gaussien.**

μ_0 est aussi la valeur plus probable de M. Dans la limite des grands nombres, les histogrammes (comme les idéogrammes) correspondants atteignent par conséquent leur maximum au voisinage de $m = \mu_0$. En outre, la distribution $\lambda(\sigma)$ est simplement, dans ce cas, la distribution marginale de $\rho(m, \sigma)$:

$$f(\sigma) = \int \rho(m, \sigma)\, dm = \lambda(\sigma) \int N(m - \mu_0, \sigma)\, dm = \lambda(\sigma)$$

de sorte que l'histogramme des erreurs donne une estimation de $\lambda(\sigma)$.

Si l'on considère maintenant les classes plus générales de statistiques gaussiennes, du type C_2 et C_3, on s'aperçoit que les raisonnements qui précèdent ne permettent pas d'obtenir à leur égard beaucoup d'informations sur la loi $\varphi(\mu)$ qui gouverne la variable $\mathcal{M}$. Il faut alors avoir recours à d'autres techniques, dites de test d'hypothèses, dont nous décrivons deux exemples (le χ^2 et le maximum de vraisemblance) par la suite.

Dans ces techniques, la question de déduire $\varphi(\mu)$ des distributions (M, Σ) est supposée résolue : en partant d'une distribution $\varphi'(\mu)$ choisie à titre d'hypothèse *(a priori)*, on cherche à évaluer le degré de compatibilité de cette hypothèse avec la distribution (M, Σ) obtenue expérimentalement.

La problématique est alors la suivante : étant donné que le résultat expérimental est ceci, quelle probabilité a-t-on pour que le résultat soit la conséquence d'un tirage selon la loi $\varphi'(\mu)$ plutôt que la conséquence d'un tirage selon $\varphi''(\mu)$? La réponse à cette question fait appel à ce que l'on appelle la *probabilité des causes*, et est fondée sur le principe de Bayes.

Théorème de Bayes

Soit A un événement physique dont la réalisation peut être la conséquence de l'une quelconque des causes (exclusives les unes des autres) $B_1, B_2, \ldots, B_n$ sans que l'on sache avec certitude laquelle de ces causes (ou hypothèses) est responsable de la matérialisation de A.

Appelons en outre $Pr(B_1), \ldots, Pr(B_n)$ les probabilités *a priori* pour que les causes $B_1, B_2, \ldots, B_n$ soient présentes indépendamment de la réalisation ou non-réalisation subséquente de A. $Pr(B_i \cap A)$ est la probabilité pour que A se produise à travers la cause B_i. $Pr(A/B_i)$ est la probabilité pour que A se produise, sachant que la condition B_i a été réalisée. $Pr(B_i/A)$ est la probabilité pour que l'antécédent B_i se soit produit, sachant que l'événement A s'est effectivement réalisé. On a :

$$Pr(B_i \cap A) = Pr(B_i) \times Pr(A/B_i)$$
$$= Pr(A) \times Pr(B_i/A)$$

par **(3)**. D'où :

(87) $$Pr(B_i/A) = \frac{Pr(B_i)}{Pr(A)}\, Pr(A/B_i)$$

Si B_k est une autre hypothèse :

(88) $$Pr(B_k/A) = \frac{Pr(B_k)}{Pr(A)}\, Pr(A/B_k)$$

en faisant le rapport de **(87)** et **(88)**

(89) $$\frac{Pr(B_i/A)}{Pr(B_k/A)} = \frac{Pr(B_i)}{Pr(B_k)}\,\frac{Pr(A/B_i)}{(Pr\, A/B_k)}$$

Si les causes B_i et B_k sont *a priori* équiprobables :

(90) $$\frac{Pr(B_i/A)}{Pr(B_k/A)} = \frac{Pr(A/B_i)}{Pr(A/B_k)}$$

En l'absence de toute information extérieure sur la probabilité *a priori* des diverses hypothèses $B_1, \ldots, B_n$, on attribuera donc à chacune d'elles, en constatant que A s'est réalisé, une probabilité *a posteriori* proportionnelle à la probabilité pour que A se produise quand celle-ci est réalisée. En termes un peu abrupts, on dira que l'on choisit des probabilités des causes proportionnelles aux probabilités des effets. Il ne faut pas oublier cependant que, si pour une raison quelconque extérieure à la réalisation actuelle de A, on ne considère pas les hypothèses $B_1, \ldots, B_n$ comme équiprobables *a priori*, c'est la formule **(89)** qu'il faut appliquer et non la formule **(90)**.

Si l'événement A ne peut survenir que par l'une quelconque des causes B_i :

(91) $$Pr(A) = \sum_{i=1}^{n} Pr(B_i \cap A)$$

par **(4)**. D'où, en portant **(91)** dans **(87)** :

(92) $$Pr(B_i/A) = \frac{Pr(B_i)\, Pr(A/B_i)}{\sum_{i=1}^{n} Pr(B_i)\, Pr(A/B_i)}.$$

Soit C un autre événement, également susceptible de se produire quand la condition B_i est réalisée. Évaluons $Pr(C/A)$, probabilité pour que C se produise, sachant que A s'est produit :

$$Pr(C/A) = \sum_{i=1}^{n} Pr(C \cap B_i/A) =$$

(93) $$\sum_{i=1}^{n} Pr(B_i/A)\, Pr(C/A \cap B_i)$$

par **(3)**, en tenant compte du fait que l'une quelconque des hypothèses B est certainement valable. Portant **(92)** dans **(93)** :

(94) $$Pr(C/A) = \frac{\sum_{i=1}^{n} Pr(B_i)\, Pr(A/B_i)\, Pr(C/A \cap B_i)}{\sum_{i=1}^{n} Pr(B_i)\, Pr(A/B_i)}$$

Illustrons **(94)** par un exemple simple. Supposons que l'on ait trois urnes. L'une contient deux pièces d'or, la seconde une pièce d'or et une pièce d'argent, la troisième deux pièces d'argent. Pierre tire au hasard une pièce de monnaie dans une urne : c'est une pièce d'or (événement A). Quelle est la probabilité pour que, en tirant la 2[e] pièce de la même urne, celle-ci soit également en or (événement C) ? Ici $N = 3$. $Pr(B_i) = \frac{1}{3}\ \forall_i$. $Pr(A/B_1) = 1$, $Pr(A/B_2) = \frac{1}{2}$. $Pr(C/A \cap B_1) = 1$. Toutes les autres probabilités sont nulles. **(94)** donne alors : $Pr(C/A) = \frac{2}{3}$.

Estimateurs

Nous avons rencontré, au cours de cette section, plusieurs exemples d'estimation de paramètres (moyennes, variances, etc.) de la densité de probabilité $\varphi(\mu)$ attachée à la variable aléatoire $\mathcal{M}$.

D'une façon générale, on appelle *estimateur* α_K^* du paramètre α associé à la loi de probabilité $\varphi(\mu, \alpha)$, le terme général d'une suite de variables aléatoires $\alpha_1^*, \ldots, \alpha_N^*$, tel que α_K^* est une fonction des K variables aléatoires $\mathcal{M}_1, \ldots, \mathcal{M}_K$, et prenant la valeur $\alpha_K^* = f(\mu_1, \ldots, \mu_K)$, considérée comme valeur approchée de α.

Par définition, on dit que α_K^* est *sans biais* si $E(\alpha_K^*) = \alpha$ et *asymptotiquement sans biais* si on a seulement $\lim_{N \to \infty} E(\alpha_N^*) = \alpha$. Il est dit *convergent en probabilité* si $\forall \eta > 0$, $\lim_{N \to \infty} Pr\{|\alpha_N^* - \alpha| \geq \eta\} = 0$, et *convergent en moyenne quadratique* si $\lim_{N \to \infty} E|\alpha_N^* - \alpha|^2 = 0$. On peut montrer qu'un estimateur convergent en moyenne quadratique est certainement convergent en probabilité. Enfin, l'estimateur est dit *convergent au sens presque sûr* si

$$\lim_{N \to \infty} \alpha_N^* = \alpha,$$

sauf pour un ensemble de probabilité nulle d'événements élémentaires possibles.

Comme la convergence en moyenne quadratique, la convergence au sens presque sûr entraîne la convergence en probabilité. Le lecteur pourra s'assurer que les estimateurs que nous avons rencontrés **(86)** sont sans biais. On peut montrer qu'ils sont convergents au sens presque sûr.

Tests d'hypothèses

Le test de χ^2

La formule **(84)** :

$$\rho(m, \sigma) = \int \varphi(\mu)\, \lambda(\sigma/\mu)\, N(m - \mu, \sigma)\, d\mu$$

relie la distribution des résultats de mesures, $\rho(m, \sigma)$, à la distribution « vraie » de la grandeur physique $\mathcal{M}$, $\varphi(\mu)$. Le passage de la première à la seconde implique, avant tout, la connaissance de la distribution des erreurs

$$\lambda(\sigma/\mu).$$

Dans un premier temps, nous allons supposer que les variables Σ et M sont indépendantes (il est toujours possible de diviser le plan (M, Σ) en régions de M, telles que, en chaque région, l'hypothèse de l'indépendance entre M et Σ soit compatible avec les données expérimentales), de sorte que

$$\lambda(\sigma/\mu) = \lambda(\sigma)$$

où $\lambda(\sigma)$ est la distribution marginale de $\rho(m, \sigma)$:

$$\lambda(\sigma) = \int \rho(m, \sigma)\, dm$$

Dans ces conditions, une évaluation correcte de $\lambda(\sigma)$ est la distribution des erreurs telle qu'elle est observée dans les N mesures (m_i, σ_i) :

$$\lambda^*(\sigma) = \frac{1}{N}\left[\sum_i \delta(\sigma - \sigma_i)\right]$$

de sorte que nous écrirons :

(95) $\rho(m, \sigma) \approx \int \varphi(\mu)\, \lambda^*(\sigma)\, N(m - \mu, \sigma)\, d\mu$

A partir d'une distribution $\varphi'(\mu)$ choisie à titre d'essai, nous pouvons alors calculer, moyennant **(95)**, la distribution $\rho'(m, \sigma)$ se rapportant à $\varphi'(\mu)$. Le test de χ^2 a pour objet de nous permettre d'évaluer le degré de compatibilité entre la répartition obtenue expérimentalement et l'hypothèse $\rho'(m, \sigma)$, en supposant que les (m_i, σ_i) obtenus sont le résultat d'un tirage de N événements avec la distribution de probabilité $\rho'(m, \sigma)$. La distribution $\varphi^*(\mu)$ qui réalise la meilleure compatibilité est acceptée comme estimation de $\varphi(\mu)$.

Distribution de χ^2

Par définition, le χ^2 pour N variables aléatoires X_i, indépendantes, est donné par l'expression :

(96) $$\chi^2 = \sum_i \chi_i^2 = \sum_i \frac{(x_i - \bar{x}_i)^2}{\sigma_i^2}$$

Dans cette expression, $\bar{x}_i$ et σ_i représentent l'espérance mathématique et la variance relative à X_i. Les X_i étant des variables aléatoires, le χ^2 est une variable aléatoire qui fluctue autour de la valeur moyenne :

(97) $$\langle\chi^2\rangle = \sum_{i=1}^{N} \frac{\langle(x_i - \bar{x}_i)^2\rangle}{\sigma_i^2} = N$$

d'après **(22)**.

La distribution de probabilité du χ^2 dépend de la forme des distributions de probabilités des X_i. Si tous les X_i suivent une distribution normale, nous avons :

$$f(x_i) = N(x_i - \bar{x}_i, \sigma_i),$$

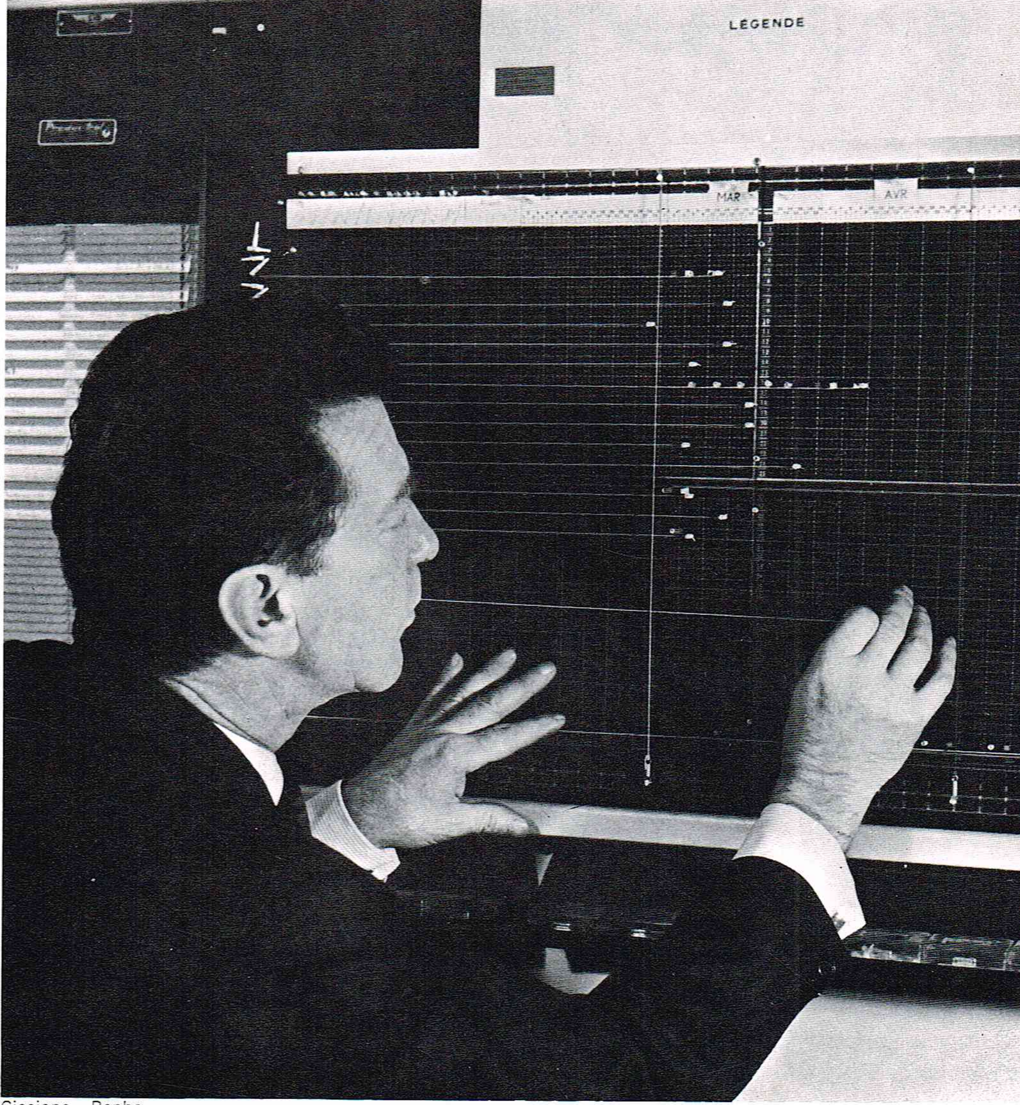
et comme les X_i sont indépendants :

$$f(x_1, \ldots, x_N) = f(x_1) \ldots f(x_N) = \frac{1}{(2\pi)^{N/2}\sigma_1 \ldots \sigma_N} \exp\left\{-\sum_{i=1}^{N} \frac{(x_i - \bar{x}_i)^2}{2\sigma_i^2}\right\}$$

La distribution de χ^2 est, d'après **(33)** :

$$F(\chi^2) = \frac{1}{(2\pi)^{N/2}\sigma_1 \ldots \sigma_N} \int \exp\left\{-\sum_{i=1}^{N} \frac{(x_i - \bar{x}_i)^2}{2\sigma_i^2}\right\} \delta\left(\chi^2 - \sum_{i=1}^{N} \frac{(x_i - \bar{x}_i)^2}{\sigma_i^2}\right) \times dx_1 \ldots dx_N$$

Posant $\xi_i = \dfrac{x_i - \bar{x}_i}{\sigma_i}$



Ciccione - Rapho

▲ ***Les lois de probabilités permettent de se faire une idée des chances que l'on a de voir survenir telle ou telle conjoncture.***

(98) $$F(\chi^2) = \frac{1}{(2\pi)^{N/2}} \int e^{-\sum_i \frac{\xi_i^2}{2}} \delta\left(\chi^2 - \sum_i \xi_i^2\right) d\xi_1 \ldots d\xi_N$$

et posant $r^2 = \sum_i \xi_i^2$

$$F(\chi^2) = \text{cte} \int_0^\infty e^{-\frac{r^2}{2}} \delta(\chi^2 - r^2)\, r^{N-1}\, dr$$

car la somme **(98)** est étendue à tout l'espace à N dimensions d'une fonction constante sur la surface de la sphère. Cette surface est proportionnelle à r^{N-1}; la constante sera identifiée ci-dessous en **(100)**. Appliquant **(31)** :

$$F(\chi^2) = \text{cte} \int e^{-\frac{r^2}{2}} \frac{\delta(\chi - r)}{2\chi} r^{N-1}\, dr$$

$$= \frac{\text{cte}}{2} e^{-\frac{\chi^2}{2}} \chi^{2\left(\frac{N-2}{2}\right)}$$

La distribution de probabilité de χ^2 est donc une distribution G :

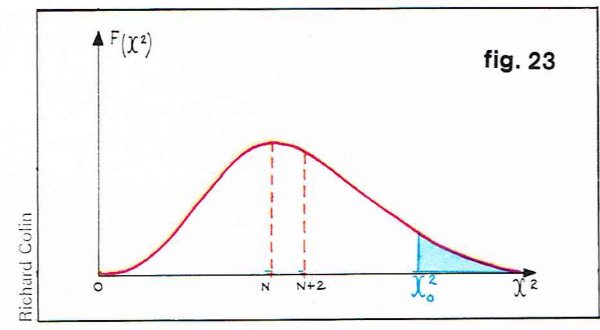
(99) $$F(\chi^2) = G\left(\frac{N-2}{2}, \frac{1}{2}, \chi^2\right)$$

(voir **27**). La constante de normalisation vaut :

(100) $$\frac{\text{cte}}{2} = \frac{\left(\frac{1}{2}\right)^{\frac{N}{2}}}{\Gamma\left(\frac{N}{2}\right)}$$

d'après **(28)**.

▶ *Figure 23 : voir développement dans le texte ci-dessous.*



Les équations **(29)** donnent alors :

$$\langle \chi^2 \rangle = N$$

en accord avec **(97)**, et

$$\sigma^2_{\chi^2} = 2\ N$$

Le χ^2 le plus probable (maximum de la densité de probabilité) est $\chi^2_m = N - 2$ (pour $N \geqslant 2$). Il ne coïncide pas avec l'espérance mathématique $\langle \chi^2 \rangle = N$.

Comportement asymptotique

Le χ^2 est une somme de variables aléatoires [les χ_i^2 de l'expression **(96)**], qui sont toutes régies par la même distribution de probabilité :

$$G\left(-\frac{1}{2}, \frac{1}{2}, \chi_i^2\right),$$

pour laquelle l'espérance mathématique ($\chi_i = 1$) et la variance ($\chi_i = 2$) existent. La seconde loi des grands nombres nous donne alors le comportement asymptotique de la distribution de χ^2 pour les grandes valeurs de N :

$$F(\chi^2) \xrightarrow[N \to \infty]{} N(\chi^2 - N, \sqrt{2N})$$

En pratique, l'expression précédente est utilisée pour $F(\chi^2)$ quand $N \geqslant 30$.

Utilisation du χ^2

Soit $\varphi_i(x)$ les distributions de probabilités se rapportant aux X_i :

$$\langle x_i \rangle = \int x\, \varphi_i(x)\, dx$$

$$\sigma_i^2 = \int (x - \langle x_i \rangle)^2\, \varphi_i(x)\, dx$$

A priori, nous pouvons prédire la probabilité pour que le χ^2 **(96)** résultant d'un tirage sur les X_i tombe dans l'intervalle $(\chi^2, \chi^2 + d\chi^2)$. Cette probabilité est $F(\chi^2)\, d\chi^2$, donnée en **(99)**. En particulier, si χ^2_0 est un nombre donné à l'avance :

$$Pr(\chi^2 \geqslant \chi^2_0) = \int_{\chi^2_0}^{\infty} F(\chi^2)\, d\chi^2 \qquad (fig.\ 23)$$

Si la limite χ^2_0 est choisie telle que, par exemple, $Pr(\chi^2 \geqslant \chi^2_0) = 0{,}001$, nous obtiendrons un $\chi^2 \geqslant \chi^2_0$ seulement une fois sur mille tirages. L'éventualité que cela se produise dans un tirage unique est pratiquement exclue.

A l'inverse, supposons les x_i donnés et les $\varphi_i(x)$ inconnus. Si, à partir d'un choix $\varphi'_i(x)$ pris à titre d'essai, nous arrivons à une valeur de χ^2 qui dépasse χ^2_0 nous conclurons qu'il est pratiquement exclu que les hypothèses $\varphi_i(x)$ soient les véritables distributions qui régissent les variables X_i. On considérera comme possibles les familles $\varphi'_i(x)$ qui mèneront à une valeur de χ^2 plus petite que χ^2_0 ; la famille $\varphi'_i(x)$ qui mènera à la valeur de χ^2 la plus petite aura notre préférence et sera retenue comme estimation des $\varphi_i(x)$.

Les familles exclues d'après le critère du χ^2 sont toujours exclues *en probabilité*. On dit que l'exclusion est faite au *degré de confiance* $1 - Pr(\chi^2 \geqslant \chi^2_0)$ (= 99,9 % dans l'exemple choisi). Le choix du degré de confiance est une question de rigueur scientifique. En général on considère comme raisonnables des niveaux de confiance de 95 ou 99 %. Il faut noter que d'un autre côté nous devons nous méfier aussi des familles $\varphi'_i(x)$ qui conduisent à des valeurs trop petites de χ^2, par exemple quand la probabilité d'obtenir un χ^2 plus grand que le χ^2 obtenu dépasse 98 %. En général, quand cela arrive, cela veut dire qu'en choisissant les $\varphi'_i(x)$ on a surestimé les variances σ_i^2 (*fig.* 24).

Application aux problèmes d'estimation

Retournons aux problèmes d'estimation et servons-nous du critère du moindre χ^2 comme critère d'estimation. Dans le cas de mesures gaussiennes répétées, les résultats de mesures se répartissent conformément à la distribution **(82)** :

$$f(m) = N(m - \mu_0, \sigma_0)$$

Définissant le χ^2 comme :

$$\chi^2 = \sum_i \frac{(m_i - \mu_0)^2}{\sigma_0^2}$$

on a pour estimation de μ_0, au sens de la méthode du moindre χ^2, la valeur μ_0^* donnée par :

$$\textbf{(101)} \qquad \left(\frac{\partial \chi^2}{\partial \mu_0}\right)_{\mu_0^*} = 0$$

$$\mu_0^* = \sum_i \frac{m_i}{N}$$

Cette valeur coïncide avec l'estimation faite en s'aidant de la loi des grands nombres.

Si les mesures ne sont pas faites avec la même précision,

$$\chi^2 = \sum_i \frac{(m_i - \mu_0)^2}{\sigma_i^2}.$$

La condition **(101)** admet pour solution :

$$\mu_0^* = \frac{\sum_i \frac{m_i}{\sigma_i^2}}{\sum_i \frac{1}{\sigma_i^2}}$$

μ_0^* est la moyenne des m_i, pondérés par les paramètres de précision $\left(\frac{1}{\sigma_i^2}\right)$.

La variable aléatoire μ_0^* a pour espérance mathématique et pour variance :

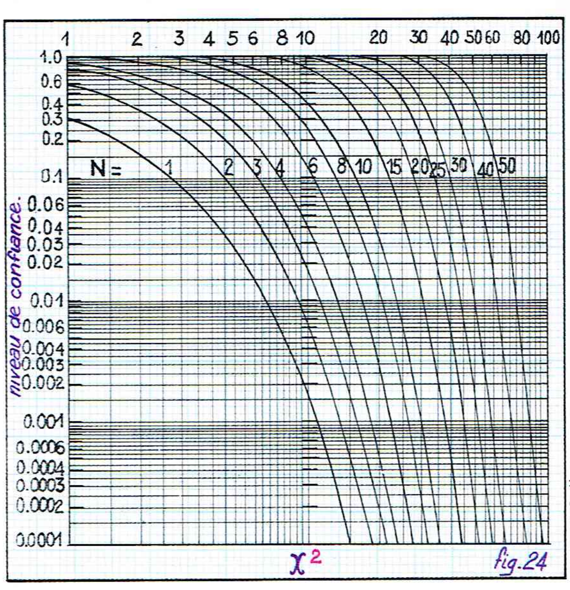
$$E(\mu_0^*) = \mu_0$$

$$\sigma^2(\mu_0^*) = \left(\sum_i \frac{1}{\sigma_i^2}\right)^{-1}$$

L'écart type $\sigma(\mu_0^*)$ nous donne la précision de l'estimation μ_0^*.

Soit N le nombre total de mesures. La relation **(101)** définit une relation entre les N variables m_i : de sorte que le χ^2 obtenu en identifiant μ_0 à μ_0^* ne doit pas se comparer à la distribution de χ^2 à N degrés de liberté, mais à la distribution de χ^2 à $N - 1$ degrés de liberté.

De façon générale, si les hypothèses $\varphi'_i(x)$ dépendent de certains paramètres inconnus α_k ($k = 1, n$), la méthode du moindre χ^2 permet d'évaluer ces paramètres à travers les n relations :



▶ *Figure 24 : le choix du degré de confiance est une question de rigueur scientifique.*

$$\left(\frac{\partial \chi^2}{\partial \alpha_k}\right)_{\alpha_k^*} = 0$$

mais le χ^2 calculé avec ces α_k^* suit alors une distribution de χ^2 à N — n degrés de liberté.

Application aux histogrammes

Le nombre N_j de mesures (m_i, σ_i) tombant dans le j-ième segment de l'histogramme, est une variable aléatoire. Si $\varphi(\mu)$ est la loi de distribution de la grandeur $\mathcal{M}$, la probabilité pour qu'une mesure tombe dans l'intervalle Δm_j, est, d'accord avec **(95)** :

$$\textbf{(102)} \qquad p_j = \int_{\Delta m_j, \sigma} \rho\,(m, \sigma)\, dm\, d\sigma$$

$$\textbf{(103)} \qquad \simeq \int_{\Delta m_j, \sigma} \varphi(\mu)\, \lambda^*(\sigma)\, N(m - \mu, \sigma)\, d\mu\, d\sigma$$

La probabilité pour qu'une mesure ne tombe pas dans le j-ième segment de l'histogramme est $q_j = 1 - p_j$. D'autre part, chaque mesure est indépendante des autres, de sorte que le nombre total N_j de mesures qui tombent dans le segment (j) de l'histogramme obéit à la distribution binominale avec les paramètres p_j et q_j donnés par **(102)** ou **(103)**. En particulier :

$$n_j = E(N_j) = Np_j$$

$$\sigma_j = \sigma(N_j) = \sqrt{Np_jq_j}$$

où N est le nombre total de mesures.

Écrivons le χ^2 :

$$\chi^2 = \sum_j \frac{(N_j - n_j)^2}{\sigma_j^2}.$$

On note que dans cette expression σ_j est la variance relative à l'espérance mathématique n_j et n'a rien à voir avec le nombre réel N_j de mesures qui tombent effectivement dans le segment considéré. En dépit de la pratique courante, il est donc incorrect de substituer σ_j par $\sqrt{N_j}$, quand on veut comparer les distributions de χ^2 avec l'expression **(99)**.

Le χ^2 obtenu nous permet d'évaluer le degré de compatibilité d'une hypothèse donnée avec l'histogramme obtenu ; l'hypothèse $\varphi^*(\mu)$ qui minimise ce χ^2 sera, éventuellement, retenue comme estimation de $\varphi(\mu)$.

Quelques réserves s'imposent cependant. On notera d'abord que, si nous avons l intervalles au total sur l'histogramme, les l variables N_j ne sont pas réellement indépendantes ; nous avons en effet la condition de « normalisation » :

$$\sum_{j=1}^{l} N_j = N$$

qui lie les N_j entre eux. Cette condition diminue le nombre de variables réellement indépendantes de l à $l - 1$, de sorte que le χ^2 suit une loi de distribution à $l - 1$ degrés de liberté.

La seconde observation est que, en toute rigueur, les distributions des N_i ne sont pas gaussiennes, mais binominales, de sorte que la distribution **(99)** n'est pas strictement valide pour le χ^2. Toutefois, nous savons que la distribution binominale tend très rapidement vers une distribution gaussienne. L'erreur commise en se servant de la distribution de χ^2 **(99)** est négligeable, pourvu que tous les $N_j \geqslant 10$ (*fig.* 10). Pour cette raison, on exclut généralement du calcul de χ^2 les cellules d'histogramme telles que $N_j < 10$.

Extension au cas de variables corrélées

Nous avons vu **(62)** qu'à partir d'un ensemble de N variables aléatoires $\underset{\sim}{X}$, centrées, gaussiennes et corrélées les unes aux autres, il est possible de trouver un ensemble de N variables aléatoires $\underset{\sim}{Y}$ gaussiennes, centrées, réduites et indépendantes. Il suffit pour cela que la matrice des variances $\underset{\sim}{V}$ **(61)** soit régulière.

Les $\underset{\sim}{Y}$ sont reliées aux $\underset{\sim}{X}$ par une transformation linéaire $\underset{\sim}{Y} = \underset{\sim}{A}\underset{\sim}{X}$. Dans cette transformation, la variable aléatoire unidimensionnelle $\chi^2 = \underset{\sim}{X}^T\underset{\sim}{V}^{-1}\underset{\sim}{X}$ est conservée :

$$\chi^2 = \underset{\sim}{Y}^T\underset{\sim}{W}^{-1}\underset{\sim}{Y} = \sum_i y_i^2 \qquad \text{car } (\underset{\sim}{A}\underset{\sim}{V}\underset{\sim}{A}^T)^{-1} = \underset{\sim}{1}.$$

Cette variable aléatoire n'est autre que le χ^2 défini en **(96)** pour les variables centrées et réduites.

Soit donc $\underset{\sim}{X}$ un ensemble de N variables aléatoires gaussiennes, admettant une matrice des variances $\underset{\sim}{V}$ régulière. Le χ^2 :

$$\chi^2 = [\underset{\sim}{X} - \bar{\underset{\sim}{X}}]^T\, \underset{\sim}{V}^{-1}\, [\underset{\sim}{X} - \bar{\underset{\sim}{X}}],$$

suit donc une loi normale de χ^2 **(99)** à N degrés de liberté.

Bien que ce résultat ne soit strictement valable que pour un ensemble de variables gaussiennes, le χ^2 défini de la même manière pour des variables aléatoires $\underset{\sim}{X}$ suivant des lois proches de la loi normale (telles les populations des segments dans les histogrammes, pourvu que $\bar{N}_j > 10$) suit, de très près, une loi de χ^2 à N degrés de liberté.

Le maximum de vraisemblance

Cette méthode de test d'hypothèse, introduite par Fischer, se base directement sur le théorème de Bayes **(89)**.

Soit $\varphi'(\mu)$ une distribution possible pour la grandeur physique $\mathcal{M}$, et $\varphi''(\mu)$ une autre distribution également possible. Soit (m_i, σ_i) la distribution expérimentale observée dans le plan (M, Σ). Le théorème de Bayes **(89)** s'écrit dans ce cas :

$$\frac{Pr\{\varphi'(\mu) / (m_i, \sigma_i)\}}{Pr\{\varphi''(\mu) / (m_i, \sigma_i)\}} = \frac{Pr\{\varphi'(\mu)\}\, Pr\{(m_i, \sigma_i) / \varphi'(\mu)\}}{Pr\{\varphi''(\mu)\}\, Pr\{(m_i, \sigma_i) / \varphi''(\mu)\}}.$$

S'il n'y a aucune raison externe de privilégier l'une plutôt que l'autre des hypothèses $\varphi'(\mu)$, $\varphi''(\mu)$, les probabilités *a priori* de ces hypothèses sont égales :

$$\textbf{(104)} \qquad \frac{Pr\{\varphi'(\mu) / (m_i, \sigma_i)\}}{Pr\{\varphi''(\mu) / (m_i, \sigma_i)\}} = \frac{Pr\{(m_i, \sigma_i) / \varphi'(\mu)\}}{Pr\{(m_i, \sigma_i) / \varphi''(\mu)\}}.$$

Les probabilités des hypothèses sont entre elles dans le même rapport que les probabilités d'obtenir la distribution (m_i, σ_i) par un tirage sur $\varphi'(\mu)$ et $\varphi''(\mu)$ respectivement.

Les épreuves (m_i, σ_i) sont le résultat d'un tirage répété N fois sur la densité de probabilité $\rho(m, \sigma)$ liée à $\varphi(\mu)$ par **(84)** ou **(95)**. On a :

$$Pr\{(m_i, \sigma_i) / \varphi(\mu)\} = Pr\{(m_i, \sigma_i) / \rho(m, \sigma)\} =$$

$$\textbf{(105)} \qquad \prod_{i=1}^{N} \rho(m_i, \sigma_i)\, dm_i\, d\sigma_i$$

par **(3)**, chaque tirage étant indépendant des autres. Supposons que les hypothèses $\varphi(\mu)$ possibles dépendent d'un paramètre inconnu, que l'on cherche à ajuster,

$$\varphi(\mu) = \varphi(\mu, \alpha), \qquad \text{alors :}$$

$$\rho(m, \sigma) = \rho_\alpha(m, \sigma).$$

L'expression **(105)**, où nous pouvons fixer arbitrairement les dm_i et les $d\sigma_i$, prend des valeurs variables selon la valeur de α. La fonction

$$F(\alpha) = A\, Pr\{(m_i, \sigma_i) / \varphi(\mu, \alpha)\},$$

où A est une constante de normalisation, peut être considérée, conformément à **(104)**, comme la distribution de probabilité de α, sachant que la séquence des N tirages sur M, Σ a donné les résultats (m_i, σ_i).

L'estimation α^* est donnée, dans la méthode du « maximum de vraisemblance », par la condition $\left(\frac{\partial F}{\partial \alpha}\right)_{\alpha^*} = 0$; cette condition peut aussi s'écrire :

$$\textbf{(106)} \qquad \left(\frac{1}{F}\frac{\partial F}{\partial \alpha}\right)_{\alpha^*} = \left(\frac{\partial}{\partial \alpha} \operatorname{Log} F\right)_{\alpha^*} = 0.$$

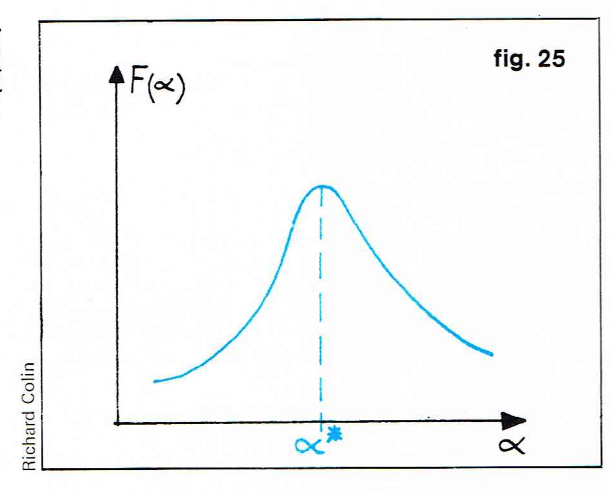
qui est plus facile à manipuler que la précédente (F n'est jamais strictement nulle, ne serait-ce que par suite des erreurs qui étendent le spectre des valeurs possibles de m de $-\infty$ à $+\infty$). Compte tenu de **(105)**, **(106)** s'écrit en effet :

$$\frac{\partial}{\partial \alpha}\left[\sum_i \operatorname{Log} \rho_\alpha(m_i, \sigma_i)\right] = 0.$$

La distribution F (α) est généralement caractérisée par son maximum α^* et par sa variance

$$\sigma_\alpha^2 = \int (\alpha - \bar{\alpha})^2\, F(\alpha)\, d\alpha \qquad (\textit{fig. } 25).$$

▶ *Figure 25 : la distribution F (α) est généralement caractérisée par son maximum α*.*



Dans l'approximation où F (α) suit une distribution normale, on a : $F(\alpha) = \frac{1}{\sigma\sqrt{2\pi}} \exp - \frac{(\alpha - \bar{\alpha})^2}{2\sigma^2}$.

La condition **(106)** donne dans ce cas $\alpha^* = \bar{\alpha}$, tandis que la variance est donnée par :

(107) $$\sigma^2_{\alpha} = -\left[\frac{\partial^2}{\partial\alpha^2} \operatorname{Log} F(\alpha)\right]^{-1}_{\alpha^*}$$

Cette relation est utilisée couramment pour définir la prévision de l'estimation α*, même quand F (α) n'est pas strictement gaussienne. On doit noter que, si F (α) est très éloignée d'une distribution normale, **(107)** perd toute signification. Seule la donnée de la distribution F (α) dans son ensemble peut alors permettre d'interpréter correctement la précision de l'estimation α*.

Relation avec la méthode du χ^2

Maximum de vraisemblance et moindre χ^2 sont deux méthodes de test d'hypothèses équivalentes pour l'estimation de la moyenne d'une variable aléatoire dont la distribution est normale. C'est en particulier le cas lorsqu'il s'agit d'estimer la valeur réelle μ_0 d'une grandeur physique $\mathcal{M}$ au moyen d'une suite de mesures gaussiennes répétées. En effet, si

$$f(m) = N(m - \mu_0, \sigma_0)$$

on a :

$$F(\mu_0) = Pr(m_1 \dots m_N/\mu_0) =$$

$$\frac{1}{(\sigma\sqrt{2\pi})^N} \exp - \frac{1}{2}\sum_i \frac{(m_i - \mu_0)^2}{\sigma_0^2} dm_1 \dots dm_N$$

$$\operatorname{Log} F(\mu_0) = -\frac{1}{2}\sum_i \frac{(m_i - \mu_0)^2}{\sigma_0^2} + \text{cte} = -\frac{\chi^2}{2} + \text{cte}$$

et : $$\left(\frac{\partial \operatorname{Log} F(\mu_0)}{\partial\mu_0}\right)_{\mu_0^*} = 0 \Leftrightarrow \left(\frac{\partial\chi^2}{\partial\mu_0}\right)_{\mu_0^*} = 0.$$

L'avantage de la méthode de maximum de vraisemblance est de permettre également d'évaluer la variance :

$$\operatorname{Log} F(\sigma_0^2) = -\frac{N}{2}\operatorname{Log}\sigma_0^2 - \frac{1}{2}\sum_i \frac{(m_i - \mu_0)^2}{\sigma_0^2} + \text{cte}$$

$$\left(\frac{\partial \operatorname{Log} F(\sigma_0^2)}{\partial\sigma_0^2}\right)_{\sigma_0^{*2}} = 0 \Rightarrow$$

$$-\frac{N}{2}\frac{1}{\sigma_0^{*2}} + \frac{1}{2}\sum_i \frac{(m_i - \mu_0)^2}{\sigma_0^{*4}} = 0$$

soit $$\sigma_0^{*2} = \sum_i \frac{(m_i - \mu_0)^2}{N}$$

On comparera ce résultat avec **(86)** où μ_0 a été remplacé par son estimation μ^*.

Exemple d'application

Considérons à titre d'exemple la détermination de la vie moyenne τ d'une substance radio-active, dans les trois situations suivantes : la durée de l'observation Δt est $\gg \tau$, et les incertitudes sur les temps de désintégration sont négligeables par rapport à τ **[a]** ; la durée de l'observation est limitée, $\Delta t < \tau$ (cas des vies moyennes longues) **[b]** ; les erreurs sur les temps de désintégration sont appréciables par rapport à τ (cas des vies moyennes très courtes) **[c]**.

a - La densité de probabilité pour qu'un atome présent au temps $t = 0$ se désintègre au temps θ est

(108) $$\varphi(\theta) = \frac{1}{\tau} e^{-\theta/\tau}.$$

Les erreurs sur les temps t_i mesurés étant négligeables, nous écrirons $\lambda(\sigma/\theta) = \delta(\sigma)$, de sorte que nous pouvons écrire :

$$f(t) = \int \rho(t, \sigma)\, d\sigma =$$

(109) $$\frac{1}{\tau}\int e^{-\theta/\tau}\, \delta(\sigma)\, N(t - \theta, \sigma)\, d\sigma\, d\theta = \frac{1}{\tau} e^{-t/\tau}$$

où nous avons utilisé **(84)**.

D'autre part :

$$F(\tau) = A\, Pr(t_1 \dots t_N/\tau) = A \sum_{i=1}^{N} \frac{1}{\tau} e^{-t_i/\tau}$$

La condition du maximum de vraisemblance **(106)** donne alors :

(110) $$\tau^* = \frac{1}{N}\sum_i t_i$$

La première loi des grands nombres assure en effet que $\lim_{N\to\infty} \tau^* = E(t) = \tau$, où E (t) est l'espérance mathématique de t, régie par la loi de probabilité **(109)**.

La précision de l'estimation τ* peut être évaluée à l'aide de **(107)** :

$$\sigma_{\tau^*} = -\left[\frac{\partial^2}{\partial\tau^2}\operatorname{Log} F(\tau)\right]^{-1/2}_{\tau^*} = \frac{\tau^*}{\sqrt{N}}$$

Bien que F (τ) ne soit pas gaussienne, cette valeur est très proche de l'écart type de la variable τ* définie par **(110)** ; en effet chacun des tirages t_i est régi par la loi **(109)** égale à $G\left(0, \frac{1}{\tau}, t\right)$ [donnée en **(27)**], dont la variance est $\sigma^2 = \tau$ d'après **(29)**. La somme des t_i admet donc pour variance $N\tau$ d'après **(35)**.

b - Si la durée d'observation Δt est limitée, on doit substituer **(108)** par :

$$\varphi(\theta) = \frac{1/\tau\, e^{-\theta/\tau}}{\int_0^{\Delta t} 1/\tau\, e^{-\theta/\tau}\, d\theta} = \frac{e^{-\theta/\tau}}{\tau(1 - e^{-\Delta t/\tau})}$$

définie sur l'intervalle $0 \leqslant \theta < \Delta t$, ce qui entraîne, au lieu de **(109)**,

$$f(t) = \frac{e^{-t/\tau}}{\tau(1 - e^{-\Delta t/\tau})}$$

$$F(\tau) = \frac{A}{\tau^N(1 - e^{-\Delta t/\tau})^N} e^{-\sum_i t_i/\tau}$$

L'estimation τ*, solution de l'équation **(106)**, est solution de :

$$-N\tau^* + \Sigma t_i + N\Delta t \frac{e^{-\Delta t/\tau^*}}{1 - e^{-\Delta t/\tau^*}} = 0$$

Soit, au premier ordre en Δt :

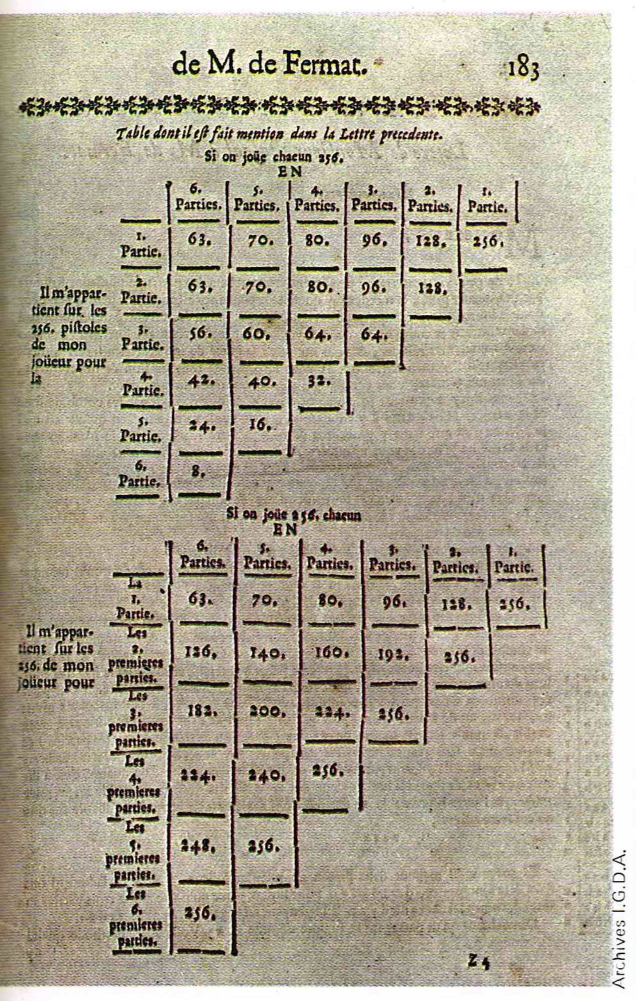
$$\tau^* \simeq \frac{1}{N}\sum_i t_i + \Delta t\, e^{-\frac{N\Delta t}{\sum_i t_i}}$$

c - Soit σ_0 la précision de mesure sur les instants des désintégrations. Portant $\lambda^*(\sigma) = \delta(\sigma - \sigma_0)$ et utilisant **(95)**, nous avons, après intégration sur σ :

$$f(t) = \frac{1}{\tau}\int e^{-\theta/\tau} N(t - \theta, \sigma_0)\, d\theta$$

D'où : $$F(\tau) = \frac{A}{\tau^N}\prod_{i=1}^{N}\int e^{-\theta/\tau} N(t_i - \theta, \sigma_0)\, d\theta$$

La solution τ* de la condition de maximum de vraisemblance **(106)** peut être recherchée par les méthodes de calcul numérique. Elles indiquent quelles corrections (petites) doivent être apportées à **(110)** pour tenir compte des erreurs de mesure.



de M. de Fermat. 183

Table dont il est fait mention dans la Lettre precedente.

Si on joüe chacun 256.
EN

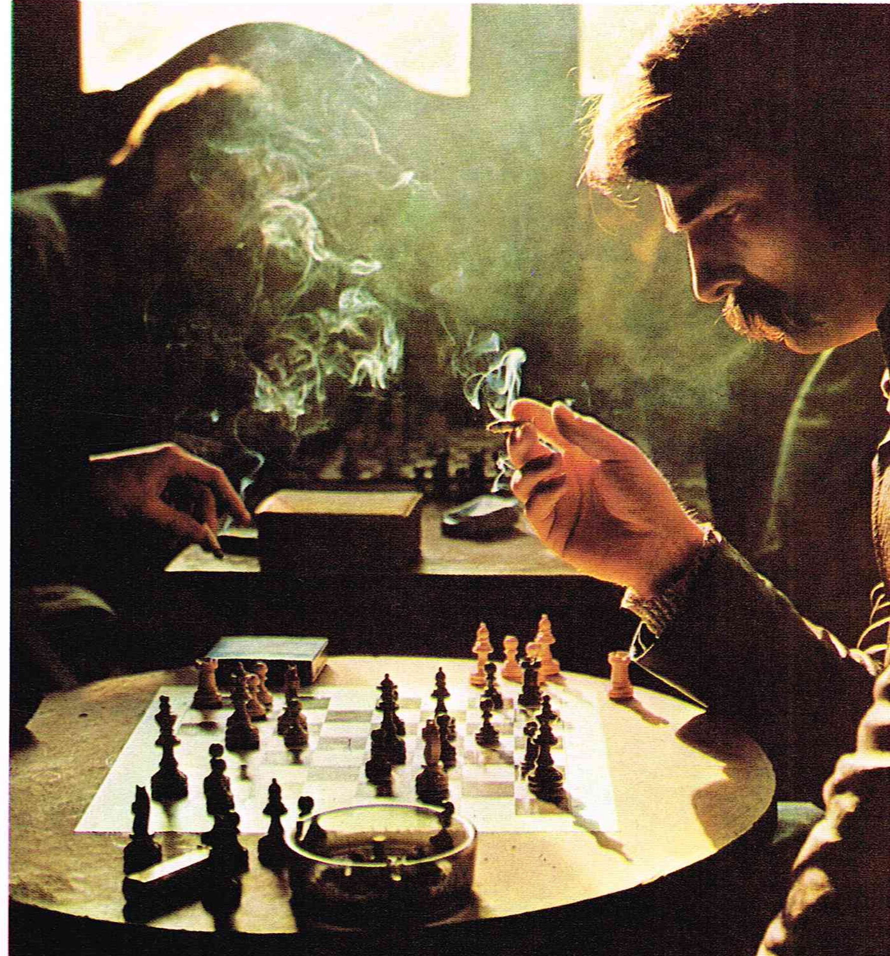
Il m'appartient sur les 256. pistoles de mon joüeur pour la		6. Parties.	5. Parties.	4. Parties.	3. Parties.	2. Parties.	1. Partie.
	1. Partie.	63.	70.	80.	96.	128.	256.
	2. Partie.	63.	70.	80.	96.	128.	
	3. Partie.	56.	60.	64.	64.		
	4. Partie.	42.	40.	32.			
	5. Partie.	24.	16.				
	6. Partie.	8.					

Si on joüe 256. chacun
EN

Il m'appartient sur les 256. de mon joüeur pour		6. Parties.	5. Parties.	4. Parties.	3. Parties.	2. Parties.	1. Partie.
	La 1. Partie.	63.	70.	80.	96.	128.	256.
	Les 2. premieres parties.	126.	140.	160.	192.	256.	
	Les 3. premieres parties.	182.	200.	224.	256.		
	Les 4. premieres parties.	224.	240.	256.			
	Les 5. premieres parties.	248.	256.				
	Les 6. premieres parties.	256.					

Z 4

Archives I.G.D.A.



A. Mendonça - Cedri

▲ *A gauche, cette page, tirée d'une édition de 1679 sur les travaux de Fermat, reproduit la table jointe à une lettre envoyée à ce même Fermat par Pascal, dans laquelle il l'entretient de questions relatives aux probabilités dans le jeu de hasard. A droite, au jeu d'échecs, la théorie ne permet pas de décider concrètement quelle est la stratégie optimale pour chacun des deux joueurs.*

Éléments de théorie des jeux

Introduction

L'analyse des jeux de hasard (pile ou face, jeu de roulette, etc.) a joué un grand rôle dans le développement de la théorie de probabilités. Blaise Pascal, qui a défini de nombreux concepts de cette dernière, avait souvent en tête les questions que lui posait son ami le chevalier de Méré, joueur invétéré. Plus tard, la théorie des probabilités a largement dépassé ce champ d'application et est devenue une science mathématique fondamentale, en particulier pour les sciences physiques.

Les relations de la théorie des jeux avec les problèmes concrets de stratégie sont plus lâches. John von Neumann, le mathématicien hongrois qui, autour de 1928, a posé les bases de cette théorie, n'avait pas pour ambition de définir la meilleure façon possible de jouer aux échecs ou au poker, mais d'analyser en termes logiques et mathématiques la situation à laquelle les joueurs d'échecs ou de poker sont confrontés, du fait de leur acceptation des règles de ces jeux. Les résultats obtenus par la théorie ne sont pas (sauf dans les cas les plus simples) directement utilisables, en ce sens qu'ils ne dictent pas la conduite que tel ou tel joueur doit tenir dans telle ou telle situation concrète. Par contre, ils permettent d'établir parfois qu'il existe une méthode optimale pour jouer à ces jeux, en fonction des buts que le joueur cherche à atteindre. Par exemple, au jeu d'échecs, la théorie permet de démontrer qu'il existe une stratégie optimale pour chacun des deux joueurs, mais elle ne permet pas de décider concrètement quelle est cette stratégie. Si elle le faisait, d'ailleurs, le jeu perdrait tout intérêt, car le résultat de la partie serait connu des deux joueurs avant même que celle-ci commence.

Cependant, de même que le champ d'application de la théorie des probabilités déborde largement les jeux de hasard, celui de la théorie des jeux déborde largement le cadre des jeux dont les résultats se comptent en points ou en argent gagnés ou perdus. En tant que théorie de l'analyse logique et mathématique des situations conflictuelles, tenant compte des résultats désirés ou redoutés par les partenaires, elle peut être utile pour éclairer les décisions des hommes politiques, des stratèges militaires, ou des hommes d'affaires placés en situation de concurrence. Ce n'est donc sans doute pas par hasard que cette branche des mathématiques a connu un vif développement après la Seconde Guerre mondiale, dans le climat de la guerre froide. Dans un tel climat, la possession des armes absolues que l'on sait faisait que les problèmes de décisions stratégiques avaient pris le pas sur les problèmes d'équipement militaire.

La théorie des jeux est donc une théorie de la *décision formelle*. Les circonstances dans lesquelles ces décisions doivent être prises peuvent englober une part de hasard (le risque), comme dans le jeu de poker, mais ce n'est pas toujours le cas. Dans de nombreuses situations, au contraire, l'effet de toutes les décisions possibles est connu avec précision, et une seule d'entre elles produit le meilleur résultat recherché. Il faut donc se garder de confondre la théorie des jeux avec celle des jeux de hasard. Cependant, comme on le verra, même dans les situations conflictuelles complètement déterminées, les joueurs ont parfois intérêt à introduire un élément d'incertitude en variant, d'une partie à l'autre et de façon aléatoire, la stratégie qu'ils choisissent. Par ce biais, la notion de probabilité est alors réintroduite dans leur jeu.

Notons enfin que, puisque la théorie prend en compte non seulement les données (règles du jeu) de la situation en question, mais aussi les résultats désirés ou redoutés par les joueurs, selon un ordre de préférence qui peut être subjectif, la théorie ne vise pas toujours à déterminer un comportement rationnel des joueurs, au sens où ce comportement serait dicté uniquement par des considérations objectives. Ce fait nous retient de parler, à propos de la théorie des jeux, de théorie de la décision rationnelle, et rapproche la théorie des jeux des sciences sociales et de la psychologie.

Utilités

Afin de théoriser les problèmes de stratégie, il est nécessaire, non seulement de bien connaître les règles du jeu, qui fixent les conséquences de chacune des décisions des joueurs et le moment où le jeu s'arrête, mais encore de chiffrer son résultat, ou, comme on dit, le règlement. Les règlements attribués en théorie des jeux à tel joueur ne correspondent pas nécessairement à des sommes perdues ou gagnées. Ce peut être $+$ **1** pour « gagné », — **1** pour « perdu » et **0** pour « match nul », mais aussi n'importe quelles autres valeurs. Il se peut, par exemple, que d'un point de vue psychologique, le joueur estime qu'il est plus dommageable pour lui de perdre qu'il n'est avantageux de gagner. Il affectera alors au règlement de la situation « perdu » une valeur — ***a***, $a > 1$, ou bien au règlement de la situation « gagné » une valeur ***b***, $b < 1$. On suppose généralement, en théorie des jeux, que les règlements sont définis sur une échelle relative, c'est-à-dire que n'importe quel changement d'échelle du type $y = ax + b$ n'affectant pas les préférences relatives des joueurs n'affecte pas les décisions de ceux-ci. Dans ce cas, on désigne les règlements sous le nom d'*utilités*. Les utilités sont donc définies à une transformation linéaire près. Il est toujours possible de choisir arbitrairement les utilités 0 et 1, et les autres utilités sont alors fixées comme dans l'exemple ci-dessus, par les appréciations relatives des autres résultats possibles du jeu.

▼ ***Figure 26 : les 23 configurations finales possibles au jeu de tic-tac-toe.***

fig. 26

Richard Colin

Naturellement, le résultat recherché par le joueur, à l'aide de la théorie des jeux, sera de *maximaliser l'utilité*. Dans les situations dans lesquelles le hasard intervient, on choisira de *maximaliser l'espérance d'utilité*, au sens de la théorie des probabilités **(19)**. Ce choix ne représente pas nécessairement un comportement rationnel (maximalisation du gain final), sauf dans l'hypothèse où les joueurs jouent un grand nombre de parties successives. Supposons, pour fixer les idées, que le joueur doive tirer un billet de loterie de l'un des deux chapeaux A et B disposés devant lui, et qui contiennent chacun 20 billets. Dans le chapeau A, 2 billets donnent droit à 100 F, 4 à 5 F, 4 sont vides, 4 donnant une amende de 5 F et 6 une amende de 10 F. Dans le chapeau B, 1 seul billet donne droit à 100 F, 6 à 5 F, 5 sont vides, 4 donnant une amende de 5 F et 4 une amende de 10 F. En termes d'espérance d'utilité, tirer un billet du chapeau A (espérance de gain de 7 F) est deux fois meilleur que tirer un billet du chapeau B (espérance de gain de 3,5 F). Cependant, en une partie, le résultat le plus probable est de perdre 10 F avec le premier billet et de gagner 5 F avec le second : un homme prudent pourrait donc préférer cette deuxième solution.

Les utilités affectées à tel ou tel résultat par chacun des joueurs ne sont pas nécessairement les mêmes, l'ordre des préférences ou les rapports entre ces préférences pouvant différer d'un joueur à l'autre. Une classe particulièrement importante de jeux, cependant, est celle des jeux à *deux personnes de somme nulle*, pour lesquelles les utilités affectées par l'un des joueurs (tels les gains et les pertes) sont égales et opposées à celles affectées par l'autre joueur. Sauf exception, nous ne traiterons dans ce qui suit que de situations rentrant dans ce cadre.

Stratégies

On appelle stratégie l'énoncé de la conduite d'un joueur, compte tenu de chacune des réactions possibles de l'adversaire à chacun des coups qu'il joue.

Le déroulement d'une partie et son utilité sont déterminés, dans un jeu à deux personnes, par un couple de stratégies : celle du joueur, et celle de l'adversaire.

Il ne faut pas confondre le nombre de stratégies et le nombre de déroulements possibles d'une partie. Considérons, pour fixer les idées, le jeu de tic-tac-toe. Ce jeu, qui se joue sur un damier 3×3 cases, consiste, pour le joueur A, à essayer d'aligner 3 croix, pendant que son adversaire B essaie, de son côté, d'aligner 3 ronds. Chacun des joueurs place alternativement une croix ou un rond, en commençant par une croix (*fig.* 26). Quand « A » a placé 5 croix, et « B » a placé 4 ronds, le damier est entièrement rempli. Nous supposerons, pour simplifier, que les joueurs décident de toujours poursuivre la partie jusqu'à compléter le damier, et ne s'arrêtent pas, comme c'est l'habitude, dès que l'un d'eux a aligné trois signes.

Il y a $\frac{9!}{5!4!} = 126$ manières distinctes de remplir le damier, et $9! = 362\,880$ déroulements possibles d'une partie.

Numérotons les cases de 1 à 9. Une stratégie possible pour A consiste à décider : « au premier coup, je place une croix dans la case 1 ; au troisième, si B a rempli la case 2, je placerai une croix en 3 ; s'il a rempli 3, je placerai une croix en 4... s'il a rempli 9, je placerai ma croix en 5. Au cinquième coup, etc. » Étant donné que B a 8 choix possibles (de 2 à 9) au deuxième coup, et que dans chacune de ces alternatives A peut décider de placer sa marque dans l'une quelconque des 7 cases libres, le nombre de stratégies possibles au troisième coup est 7^8. De même, au cinquième, il est de 5^6, et au septième, de 3^4. Au neuvième coup, il n'y a qu'une stratégie possible, celle de placer sa croix dans la seule case restée libre. Au total, le nombre de stratégies possibles pour l'ensemble de la partie est :

$$9 \times 7^8 \times 5^6 \times 3^4 \simeq 6{,}6\ 10^{13}.$$

Tous ces nombres sont, en réalité, surévalués car il est raisonnable de considérer comme équivalentes toutes les figures qui ne diffèrent que par une symétrie par rapport à un bord du damier ou par une rotation de 90°. Ainsi, compte tenu de ces symétries, le nombre de manières distinctes de

remplir le damier tombe de 126 à 23, et le nombre de déroulements possibles tombe de 362 880 à

$$23 \times 5! \times 4! = 66\,240.$$

Le nombre de stratégies possibles est donc, en général, très grand, même pour les jeux les plus simples. Si grand que soit ce nombre, cependant, il ne peut être infini, à partir du moment où le nombre de coups possibles et le nombre de choix à chaque coup restent finis.

Dès que chacun des joueurs a choisi une stratégie, le déroulement de la partie devient inéluctable. Si le joueur ignore la stratégie choisie par son adversaire, il ne peut prédire le résultat de la partie, mais un observateur qui saurait quel couple de stratégies a été choisi pourrait le faire. On peut donc associer à tout couple de stratégies (i, j) un résultat, ou une utilité, que nous désignerons par U_{ij}. La matrice U_{ij}, où i varie de 1 à n (nombre de stratégies possibles pour le premier joueur) et j de 1 à m (nombre de stratégies possibles pour le deuxième), est appelée *matrice de jeu*.

Stratégies dominantes et minimax

Considérons, dans un jeu à deux personnes de somme nulle, l'ensemble des stratégies possibles A_i pour le joueur A et l'ensemble des stratégies possibles B_j pour B. Soit U_{ij} la matrice de jeu.

Comparons deux stratégies possibles pour A, soit A_k et A_l. Si les utilités procurées par A_k sont supérieures ou au moins égales à celles procurées par A_l, quelle que soit la stratégie B_j choisie par B, on dit que A_k *domine* A_l. Si A_k domine toutes les autres stratégies $A_l \neq A_k$, A_k est une *stratégie dominante*.

On peut distinguer des jeux pour lesquels :

— il existe une stratégie dominante pour chacun des joueurs ;

— il n'existe de stratégie dominante que pour l'un des joueurs ;

— il n'existe pas de stratégie dominante.

Dans le premier cas, la conduite à tenir pour les joueurs A et B est évidente : ils choisiront chacun la stratégie dominante correspondante. Le résultat du jeu est déterminé par ce choix.

Dans les deux autres cas, le choix des joueurs n'est pas évident.

Cependant, chaque stratégie A_i offre des utilités U_{ij} variables selon la stratégie choisie par l'adversaire. Considérons la plus mauvaise d'entre elles :

$$U_i^* = \text{Min}\ (U_{i1}, ..., U_{in}).$$

En l'absence de stratégie dominante, un joueur avisé A (et qui suppose que son adversaire saura profiter de toute imprudence de sa part) choisira la stratégie A_k qui assure le meilleur résultat parmi les plus mauvais $U^*_k = \text{Max}\ (U^*_1, ..., U^*_m)$. Tout autre choix laisserait à B la possibilité de lui infliger un résultat plus défavorable. Son adversaire, B, appliquera naturellement la même tactique. Finalement, les joueurs A et B choisiront un couple de stratégies correspondant à un *col* (le col est un point le plus bas dans une ligne horizontale de la matrice de jeu, et un point le plus haut dans la ligne verticale correspondante). On l'appelle aussi un *minimax*.

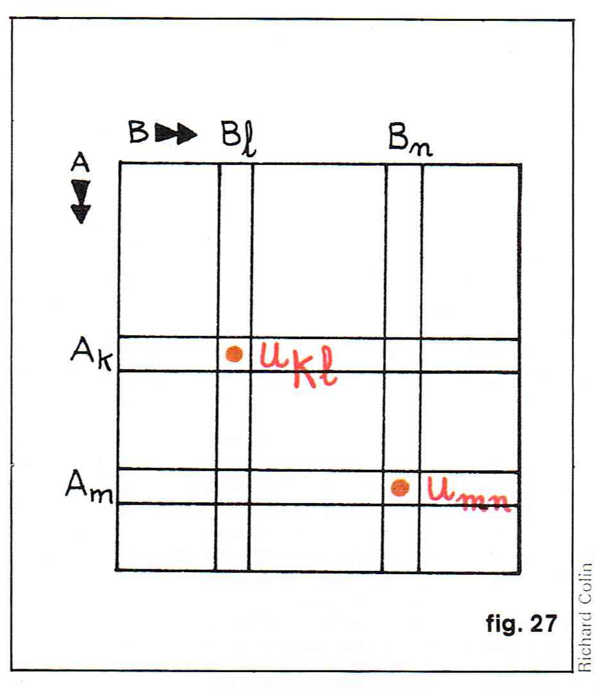
Si U_{kl} et U_{mn} sont deux cols, alors ils ont même valeur, et U_{kn} et U_{ml} sont également deux cols de même valeur : n'importe lequel de ces couples de stratégies produit donc la même utilité. En effet, par définition des cols, on peut écrire :

$$U_{kl} \geqslant U_{ml} \geqslant U_{mn}$$
$$U_{kl} \leqslant U_{kn} \leqslant U_{mn}$$

D'où $U_{kl} = U_{ml} = U_{mn} = U_{kn}$ (*fig.* 27).

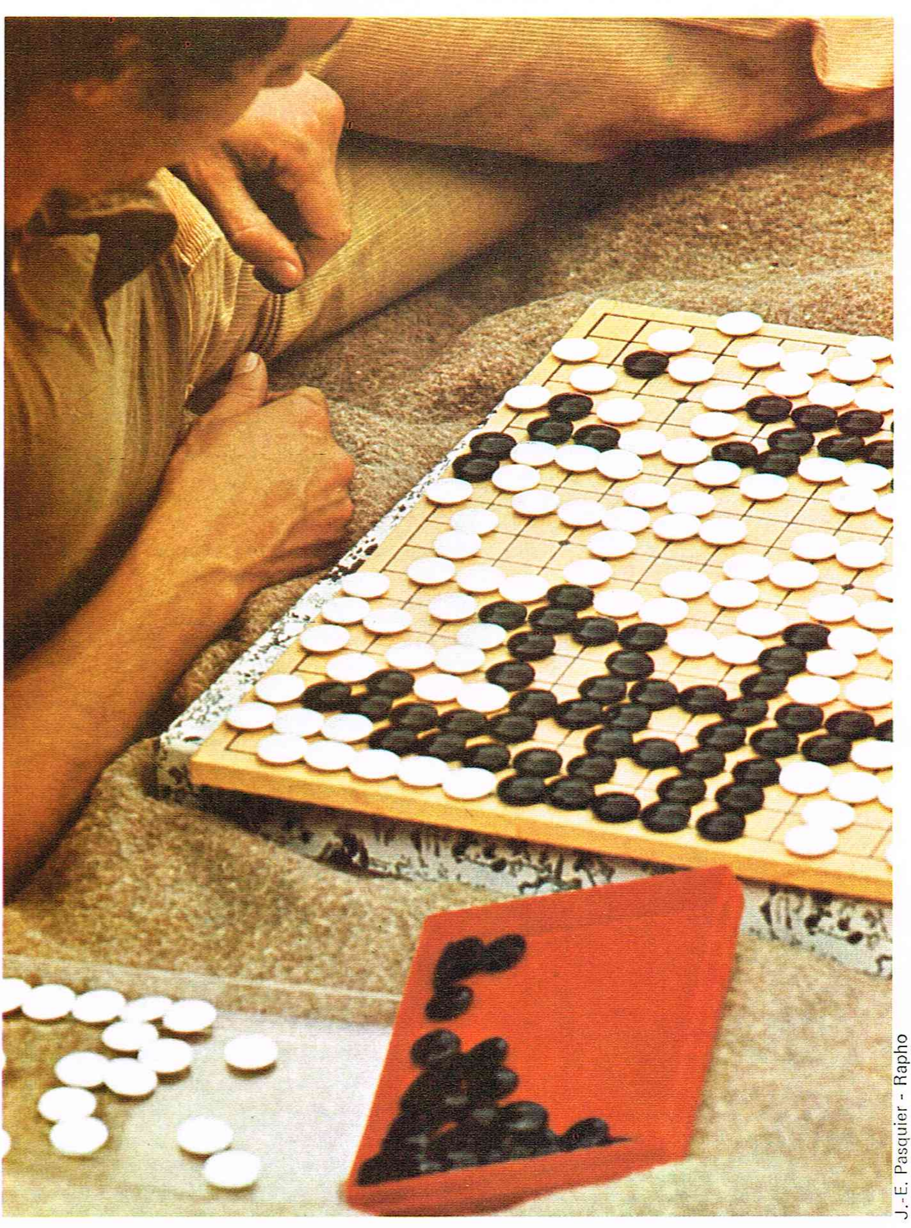
Le *principe minimax* selon lequel les joueurs doivent choisir un col de la matrice de jeu, s'il en existe, constitue donc la règle à suivre pour tous les jeux à deux personnes de somme nulle.

Tous les jeux *à information parfaite* (c'est-à-dire dans lesquels les décisions prises par l'adversaire au coup précédent sont connues du joueur, et pour lesquels, par conséquent, le hasard ne joue aucun rôle) présentent des cols. En effet, considérons les deux derniers coups joués, et supposons, pour fixer les idées, que ceux-ci sont joués successivement par A (n alternatives) et B (m alternatives). La matrice de jeu correspondant à ce sous-jeu à deux coups possède n lignes et m^n colonnes. Parmi



◄ *Figure 27 : voir développement dans le texte ci-contre.*

Richard Colin



▼ *Le jeu de* Go. *Tous les jeux à information parfaite, présentent des cols.*

J.-E. Pasquier - Rapho

ces m^n colonnes, il en existe toujours au moins une qui regroupe tous les minimums correspondant aux n lignes. La stratégie correspondante peut être en effet reformulée ainsi : choisir l'alternative qui conduit au minimum de gain pour A. Par définition, le maximum de ces minimums est un col. Il est donc superflu de jouer ces derniers coups, dont le résultat, pour des joueurs avisés, est connu d'avance. Le jeu se trouve donc ramené à un jeu amputé de deux coups, et ainsi de suite.

Dans la pratique, les dimensions des matrices correspondant aux jeux réels sont le plus souvent telles qu'il est impossible d'expliciter les stratégies qui conduisent à des cols. C'est ce qui permet aux jeux à information parfaite, tels que le jeu d'échecs, de garder un intérêt car, dans le cas contraire, toutes les parties de jeux d'échecs disputées par des joueurs avertis aboutiraient au même résultat (ou à des résultats équivalents en termes d'utilités).

Stratégies mixtes

Les jeux ou les situations conflictuelles à information parfaite sont exceptionnels. Dans la plupart des situations concrètes, les adversaires ont la possibilité de cacher l'un à l'autre une partie des décisions qu'ils prennent au cours de la partie. Les décisions tenues secrètes par le joueur B, par exemple, empêchent le joueur A d'envisager un comportement différent selon ces décisions. Par rapport au jeu à information parfaite où toutes les décisions sont connues, le nombre de stratégies possibles pour A s'en trouve diminué, et la matrice de jeu réduite. Cette matrice réduite peut encore, éventuellement, présenter un col, mais il arrive fréquemment qu'elle n'en présente plus.

Illustrons ces considérations par un exemple simple. Considérons le jeu suivant : A met dans sa main ouverte (droite ou gauche) un bouton, puis B fait de même. Les joueurs comparent alors leurs mains, et le règlement s'établit selon le schéma suivant (*fig.* 28) :

A a le choix entre deux stratégies,
« D » et « G » (choisir la main droite ou la main gauche).
B a le choix en quatre stratégies :

1 choisir *d* quel que soit le choix de A
2 choisir *d* si D et *g* si G
3 choisir *g* si D et *d* si G
4 choisir *g* quel que soit le choix de A.

La matrice de jeu correspondante est donc :

	1	2	3	4
D	2	2	— 1	— 1
G	— 2	3	— 2	3

Elle comporte, comme prévu, un col en D3, de valeur — 1. Le joueur A doit mettre le bouton dans la main droite (pour ne pas risquer de perdre 2 points) et B doit le mettre dans la main gauche pour gagner 1 point.

Ce jeu est évidemment dépourvu d'intérêt. Examinons donc la variante selon laquelle A, au lieu de mettre le bouton dans sa main ouverte, garde sa main fermée et ne l'ouvre qu'une fois que B a fait son propre choix. Les stratégies possibles pour A restent les mêmes, mais les stratégies 2 et 3, pour B, n'ont plus de sens : les seules possibilités qui s'offrent à lui sont de choisir lui-même soit de placer le bouton dans la main droite, soit dans la gauche. La matrice 2×2 correspondant à ce nouveau jeu est :

	1	4
D	2	— 1
G	— 2	3

Si A joue une seule partie, son comportement dépendra essentiellement de son caractère. S'il est audacieux, il cachera le bouton dans la main gauche, espérant que B fera de même. S'il est prudent, il préférera le cacher dans la main droite, pour ne pas risquer de perdre plus d'un point.

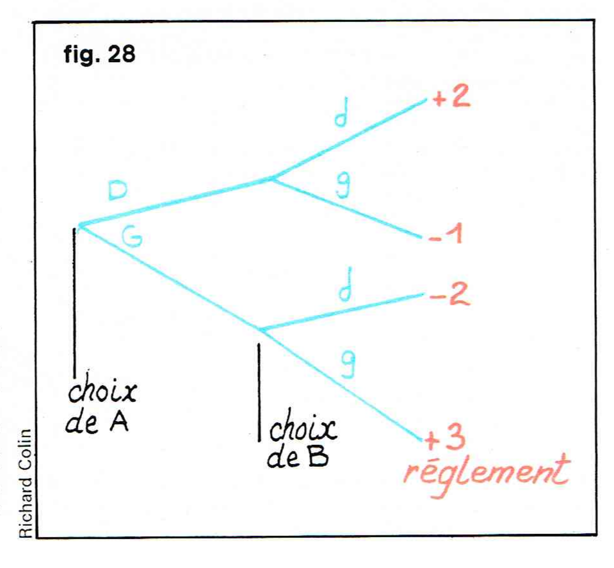
Dans un jeu répété un grand nombre de fois, A ne peut pas adopter toujours la même stratégie. S'il le faisait, B aurait tôt fait de le remarquer et d'adapter sa propre stratégie de façon à gagner. Ainsi, si A choisissait de toujours cacher le bouton dans la main gauche, B choisirait toujours la droite et A perdrait 2 points à chaque partie. Pour confondre son adversaire, A doit donc choisir tantôt une stratégie, tantôt une autre ; et pour être sûr que B ne risque pas de deviner son choix, il changera de stratégie de façon aléatoire, tirant au sort entre la stratégie D, à laquelle il affectera la probabilité x, et la stratégie G (probabilité $1 - x$). Bien entendu, B fera de même, affectant la probabilité y à la stratégie 1 et $1 - y$ à la stratégie 4.

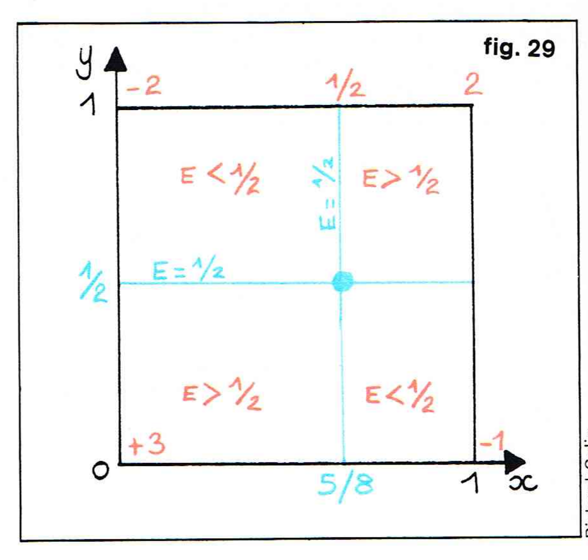
L'espérance d'utilité, dans ce cas, est :

$$E = 2\,xy - x\,(1 - y) - 2\,y\,(1 - x) + 3\,(1 - x)\,(1 - y) = y\,(8\,x - 5) - 4\,x + 3.$$

Pour $x = \frac{5}{8}$, la fréquence y choisie par B n'affecte pas l'espérance d'utilité, $E = \frac{1}{2}$. En choisissant cette valeur de x, A est donc assuré, à la longue, d'obtenir un demi-point par partie (en moyenne).

D'autre part, $E = x\,(8\,y - 4) - 5\,y + 3$. Pour $y = \frac{1}{2}$, la fréquence choisie par A n'affecte pas non plus l'espérance d'utilité, $E = \frac{1}{2}$. Il est facile de montrer que le couple de fréquences $\left(x = \frac{5}{8}, y = \frac{1}{2}\right)$ est celui qui conduit au meilleur résultat possible à la fois pour A et B. Par exemple, si A choisit $x > \frac{5}{8}$, B peut réagir en choisissant $y < \frac{1}{2}$, ce qui assure une espérance d'utilité inférieure à $\frac{1}{2}$ (*fig.* 29).





► *A gauche, figure 28 : représentation du règlement des joueurs au bouton (voir texte ci-dessus). A droite, figure 29 : l'existence d'une stratégie mixte optimale est une propriété générale des jeux à deux personnes de somme nulle.*

Le couple de fréquences $\left(x = \frac{5}{8}, y = \frac{1}{2}\right)$ est appelé, dans ce cas, la *stratégie mixte optimale*, et l'espérance de gain associée, $E = \frac{1}{2}$, est appelée *valeur du jeu*. L'existence d'une stratégie mixte optimale (ou de plusieurs stratégies mixtes optimales de même valeur) est une propriété générale des jeux à deux personnes de somme nulle. Si la matrice de jeu possède un col, ce col est un cas particulier de stratégie optimale (pour lequel les fréquences affectées sont 1 et 1). Comme dans le cas des cols, cependant, il est le plus souvent impossible de résoudre en pratique le problème de déterminer quelle est ou quelles sont la ou les stratégies mixtes optimales correspondant à une situation concrète.

Examinons les conditions que la stratégie mixte optimale d'un jeu à deux personnes de somme nulle doit remplir. Appelons U_{ij} les éléments de la matrice de jeu à n lignes (stratégies $A_1 \dots A_n$) et m colonnes (stratégies $B_1 \dots B_m$). Les probabilités affectées par B aux stratégies $A_1 \dots A_n$ sont $x_1, \dots, x_n$, et les probabilités affectées par B aux stratégies $B_1 \dots B_m$ sont $y_1, \dots, y_m$. L'espérance d'utilité est :

$$E = \sum_{i=1}^{n} \sum_{j=1}^{m} U_{ij}\, x_i\, y_j.$$

Quelle que soit la stratégie choisie par B, l'espérance de gain de A doit être au moins égale à la valeur du jeu, soit v. Par conséquent, pour chaque m :

(111) $$\sum_{i=1}^{n} U_{ij}\, x_i \geqslant v \quad (m \text{ inéquations}).$$

D'autre part, les x_i, étant des probabilités, doivent satisfaire les relations :

$$x_i \geqslant 0 \quad (n \text{ inéquations})$$

(112) $$\sum_{i=1}^{n} x_i = 1$$

De même, les relations auxquelles doivent satisfaire les y sont :

(113) $$\sum_{j=1}^{m} U_{ij}\, y_j \leqslant v \quad (n \text{ inéquations})$$

équivalentes à $$\sum_{j=1}^{m} - U_{ij}\, y_j \geqslant - v$$

$$y_j \geqslant 0 \quad (m \text{ inéquations})$$

(114) et $$\sum_{i=1}^{m} y_j = 1.$$

La résolution des systèmes d'inéquations **(111)** à **(114)** peut se faire par approximations successives.

Notons enfin que, dans le cas simple où $n = m = 2$, la résolution du problème de stratégie optimale est immédiate, en utilisant la propriété, remarquée dans l'exemple du jeu de bouton-bouton, que, au point x^*, y^* cherché, $\overrightarrow{\text{grad}}\, E$ est nul (l'espérance d'utilité ne dépend ni de la valeur de x, pour $y = y^*$ fixé, ni de la valeur de y, pour $x = x^*$ fixé).

Ce point est donc solution de $\frac{\partial E}{\partial x} = \frac{\partial E}{\partial y} = 0$.

Si nous écrivons la matrice de jeu sous la forme :

(115)

		B_1	B_2
	Pr	y	$1 - y$
A_1	x	a	b
A_2	$1 - x$	c	d

l'espérance d'utilité est :

(116) $$E = axy + bx\,(1 - y) + cy\,(1 - x) + d\,(1 - x)\,(1 - y)$$

et : $$\frac{\partial E}{\partial x} = \frac{\partial E}{\partial y} = 0 \Rightarrow$$

(117) $$x^* = \frac{d - c}{(a + d) - (b + c)} \qquad y^* = \frac{d - b}{(a + d) - (b + c)}$$

On pourra vérifier sans difficulté que, si la matrice ne possède pas de col, le point (x^*, y^*) se trouve bien dans le carré $0 \leqslant x \leqslant 1$ et $0 \leqslant y \leqslant 1$.

Nous avons considéré jusqu'ici des jeux à deux personnes de somme nulle dont les joueurs, connaissant parfaitement toutes les règles du jeu, peuvent construire, au moins en principe, la matrice de jeu. L'examen de celle-ci dicte alors leur conduite. Cependant, dans les situations de conflit réel, il n'est pas rare que les règles du jeu échappent, au moins en partie, aux joueurs, qui ne peuvent apprécier qu'après coup le résultat global des stratégies qu'ils choisissent et comparer leurs différentes efficacités — c'est-à-dire leurs différentes utilités — au terme d'une longue suite ou répétition de parties.

Pour examiner quel peut être, dans cette situation, le comportement des joueurs, reprenons l'exemple simple du jeu à deux personnes de somme nulle qui ne présente que deux stratégies possibles, à la fois pour A et B. La matrice de jeu est de la forme **(115)**. Si elle présente une stratégie dominante ou un col, un examen empirique des résultats acquis au cours d'un nombre suffisant de parties aura tôt fait de convaincre les joueurs de l'existence de cette stratégie dominante ou de ce col. Dans ce cas, en effet, l'attitude des joueurs est déterminée sans ambiguïté : ils doivent appliquer les stratégies correspondantes, qui leur assurent le maximum de gain sans risque.

Dans le cas d'un jeu dont la matrice ne présente pas de col, toutefois, la situation est différente : les joueurs doivent appliquer une stratégie mixte dont les probabilités associées, x^* et y^*, ne peuvent être évaluées que s'ils connaissent la matrice dans son ensemble. Au cours des premières parties, ils se fixent donc *a priori* une stratégie mixte caractérisée par les probabilités x et y. Celle-ci leur assure, en moyenne, une utilité donnée par **(116)**. Au bout d'un certain nombre de parties, le joueur A peut essayer d'améliorer son gain en modifiant la probabilité x associée à la stratégie A_1. Si l'utilité moyenne obtenue avec cette nouvelle stratégie mixte est plus élevée, il en déduit logiquement qu'il a eu raison de modifier, dans cette direction, la probabilité x et la modifiera encore, jusqu'à optimisation de son gain.

Naturellement, pendant ce temps, le joueur B fera de même. De plus, il est naturel de supposer que chacun des joueurs changera d'autant plus souvent, ou d'autant plus fortement, la probabilité x (pour A) ou y (pour B) que l'amélioration du gain obtenu lors de la dernière série aura été plus forte. Au total, leur comportement sera conforme à ce que l'on appelle un *modèle dynamique de jeu*, que l'on peut caractériser par le système d'équations :

(118) $$\frac{dx}{dt} = \alpha^2 \frac{\partial E}{\partial x} \qquad \frac{dy}{dt} = -\beta^2 \frac{\partial E}{\partial y}$$

où les constantes positives α^2 et β^2 mesurent l'intensité avec laquelle les joueurs A et B réagissent aux améliorations obtenues en variant les paramètres des stratégies mixtes qu'ils appliquent. La 2e équation **(118)** comporte un signe —, du fait que si le gain de A est E, le gain de B est — E.

Explicitant ces équations à l'aide de **(116)**, le système **(118)** se réduit au système d'équations :

$$\frac{dx}{dt} = uy + v \qquad \frac{dy}{dt} = -\eta x - \theta$$

où $u = \alpha^2\,(a + d - b - c)$, $v = \alpha^2\,(b - d)$, $\eta = \beta^2\,(a + d - b - c)$, $\theta = \beta^2\,(c - d)$.

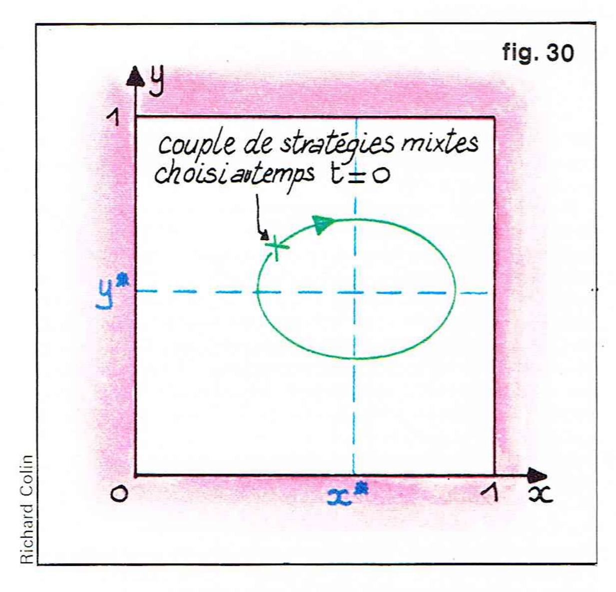
Ce système admet pour solution :

$$x = A \cos \omega t + B \sin \omega t + x^*$$

$$y = \frac{\beta}{\alpha}\left[B \cos \omega t - A \sin \omega t\right] + y^*$$

où les x^*, y^* sont donnés par **(117)** et $\omega = \sqrt{u\eta}$. Les partenaires choisissent des stratégies mixtes qui, au cours du temps, décrivent dans le plan (x, y) une ellipse centrée

▶ *Figure 30 : les partenaires choisissent des stratégies mixtes qui, au cours du temps, décrivent une ellipse centrée sur x* et y*, couple de stratégies mixtes optimal.*



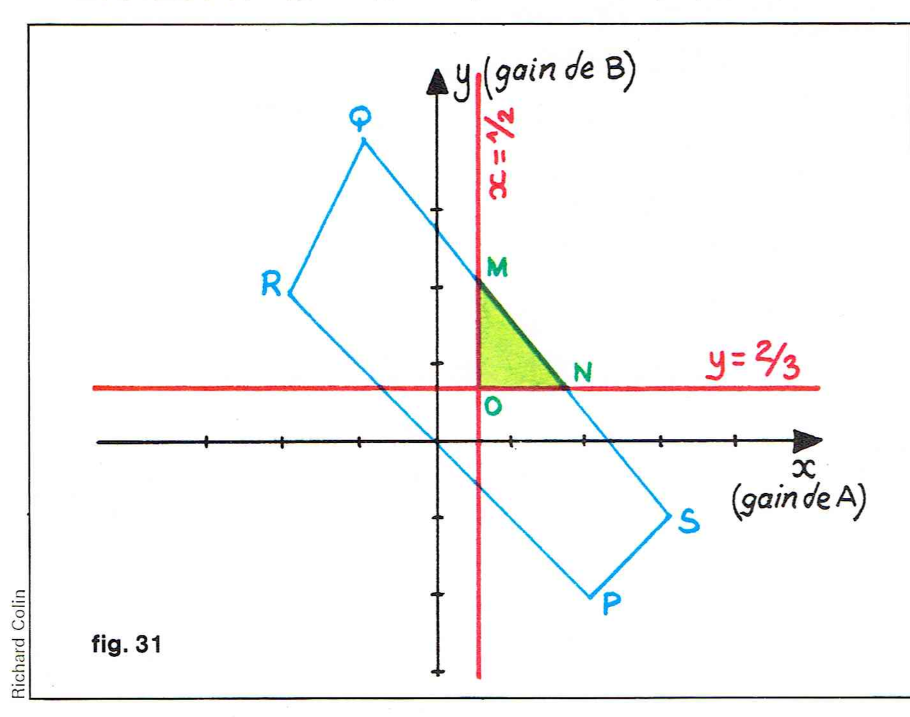
sur x^* et y^*, couple de stratégies mixtes optimal (*fig.* 30). Cette ellipse est décrite avec une vitesse angulaire uniforme. Si le point initial n'est pas trop éloigné du point x^*, y^*, de telle sorte que l'ellipse est décrite dans son entier sans sortir du carré $0 \leqslant x \leqslant 1$, $0 \leqslant y \leqslant 1$, le gain moyen obtenu par les deux joueurs au long d'un cycle complet sera le même que s'ils avaient choisi dès l'abord le couple optimal x^*, y^*.

Jeux de somme non nulle ; ententes

Pour conclure ce rapide examen des principes fondamentaux de la théorie des jeux, examinons quelques cas de jeux à deux personnes de somme non nulle, c'est-à-dire tels que les utilités attribuées à chaque résultat par les deux joueurs ne sont pas exactement opposées. Ce sera le cas, par exemple, si les préférences des joueurs ne sont pas rangées par chacun d'eux dans le même ordre avec les mêmes poids relatifs, ou si le règlement attribue des gains différents à chaque joueur, par le truchement d'une « banque », de telle sorte que ce qui est gagné par A n'est pas nécessairement perdu par B.

Dans ce cas nous n'avons plus affaire à une matrice de jeu, mais à deux : la matrice des utilités évaluée du point de vue de A, et celle évaluée du point de vue de B. L'attitude face à ces jeux sera illustrée par un exemple simple. Supposons que la matrice de jeu pour le joueur A soit

▼ *Figure 31 : diagramme du jeu avec entente (voir développement dans le texte ci-contre).*



la matrice U_1, tandis que pour le joueur B elle est U_2 :

	B_1	B_2
A_1	2	— 1
A_2	— 2	3

U_1

	B_1	B_2
A_1	— 2	4
A_2	2	— 1

U_2

Si le jeu était de somme nulle, A aurait intérêt à appliquer la stratégie mixte caractérisée par la fréquence $x^* = \frac{5}{8}$ solution de la première des équations **(117).** De même, B aurait intérêt à appliquer la stratégie mixte caractérisée par la fréquence $y^* = \frac{5}{9}$. Dans le cas où A et B appliquent ce couple de stratégie mixte, l'espérance de gain pour A est :

$$E_A = \frac{25}{72}(+2) + \frac{20}{72}(-1) + \frac{15}{72}(-2) + \frac{12}{72}(+3) = \frac{1}{2}$$

tandis que pour B elle est :

$$E_B = \frac{25}{72}(-2) + \frac{20}{72}(+4) + \frac{15}{72}(+2) + \frac{12}{72}(-1) = \frac{2}{3}$$

Cependant, il est facile de se rendre compte que les joueurs A et B ont intérêt à s'entendre.

Supposons, par exemple, qu'ils décident de partager leurs gains. Dans ce cas, leur entente qui est totale a pour effet de remplacer chacune des matrices de jeu U_1 et U_2 par la matrice $U_3 = \frac{(U_1 + U_2)}{2}$:

	B_1	B_2
A_1	0	$\frac{3}{2}$
A_2	0	1

U_3

telle que A choisit certainement la stratégie dominante A_1 et B la stratégie dominante B_2. Le couple (A_1, B_2) assure à chacun un gain de $\frac{3}{2}$ par partie, soit une amélioration de gain de 1 point pour A et de $\frac{5}{6}$ de point pour B par rapport à ce qu'ils pourraient attendre d'un jeu sans entente.

Naturellement, B pourra faire valoir que ce partage des bénéfices de l'entente est injuste, et réclamer l'application de matrices de règlement légèrement différentes, telles que $U_4 = U_3 - \frac{1}{12}$ et $U_5 = U_3 + \frac{1}{12}$ (pour les joueurs A et B respectivement), qui assurent une équipartition des améliorations et gains. Leur jeu, cependant, n'en sera pas modifié.

L'entente entre A et B, cependant, peut encore améliorer les gains de chacun des joueurs, même si elle ne remet pas en cause le mode d'attribution des règlements. Pour le voir, portons en abscisse du diagramme (*fig.* 31) le gain de A, et en ordonnées le gain de B. Les couples de stratégies A_1B_1, A_1B_2, A_2B_1 et A_2B_2 déterminent les points P, Q, R, S. Toute entente entre les joueurs ne peut que conduire à un point intérieur du quadrilatère PQRS. Cependant, en l'absence d'entente, A est assuré d'obtenir au moins $x = \frac{1}{2}$, et y est assuré d'obtenir $\frac{2}{3}$. Ils n'accepteront donc de s'entendre que pour un couple de stratégies mixtes coordonnées qui améliore leurs certitudes de gain, c'est-à-dire pour un point quelconque du triangle MNO. Ils préféreront, en outre, arrêter leur choix sur un point du segment MN, car seuls les points de ce segment ne leur permettent pas d'améliorer simultanément leurs gains par un déplacement dans OMN. Le segment MN constitue l'*ensemble d'entente*. La discussion entre A et B portera donc seulement sur la question de savoir lequel des points de ce segment sera retenu. Plus le point retenu sera proche de M, meilleur sera l'avantage retiré par B. Plus il sera proche de N, meilleur sera l'avantage retiré par A. Le choix final dépend donc de l'influence respective de A et B et de l'idée qu'ils se font d'un juste règlement.

ANALYSE DES DONNÉES

L'analyse des données regroupe les méthodes utilisées pour « décrire » de vastes ensembles de données. La complexité des phénomènes naturels se laissant rarement appréhender par l'observation d'une ou deux variables, l'intérêt d'une étude multidimensionnelle est d'éviter des choix *a priori* de caractères significatifs.

Seul l'ordinateur peut traiter des tableaux de grande dimension, ce qui explique le développement récent de telles méthodes. Nous ne parlerons pas ici de la phase préliminaire à toute analyse statistique : celle de la collecte des données, qui, ainsi que l'interprétation des résultats obtenus, ne concerne pas seulement le statisticien.

Il existe plusieurs techniques d'analyse multidimensionnelle, toutes ayant pour objectif de discerner les traits principaux des données recueillies.

Analyse factorielle

La plus ancienne de toutes les analyses factorielles, l'**analyse en facteurs communs et spécifiques** (Spearman), a été élaborée pour résoudre des problèmes du type suivant : à partir de relevés de notes obtenues par des enfants (les « *individus* ») à un certain nombre d'épreuves scolaires (les « *variables* » $X_1, X_2..., X_j, ...X_p$), il s'agit de trouver un *petit nombre* de nouvelles variables $F_1, F_2,..., F_l, ...F_k$, les *facteurs communs et spécifiques*, caractérisant le niveau scolaire d'un enfant.

On suppose que chaque X_j s'exprime comme combinaison linéaire des facteurs F_l ; l'ensemble des notes mesurées par les X_j peut ainsi être remplacé, de manière approximativement équivalente, par un ensemble de notes mesurées par les F_l ; les facteurs F_l « expliquent » donc la réussite — ou l'échec — scolaire d'un enfant.

On voit que cette analyse exige des hypothèses restrictives, contrairement aux autres analyses factorielles qui vont être exposées.

Analyse factorielle en composantes principales

Supposons qu'on dispose pour classer un ensemble d'individus de chacune de leurs mesures sur un ensemble de caractères quantitatifs appelés variables. On peut ranger ces observations dans un tableau X à n lignes et p colonnes si on suppose que les individus sont au nombre de n et les variables au nombre de p ; x_{ij} représente dans ce tableau la valeur de la j-ième variable x_j pour le i-ième individu.

Dans un langage géométrique, on peut représenter le tableau X de deux façons : soit n points dans $\mathbb{R}^p$, le i-ième point M_i représentant le i-ième individu et ayant pour coordonnées sur le j-ième axe cartésien de $\mathbb{R}^p$ (associé à X_j) la valeur prise par cet individu pour la j-ième variable (dans le cas de trois variables, voir *fig.* 1) ; soit p points (caractères) dans l'espace $\mathbb{R}^n$.

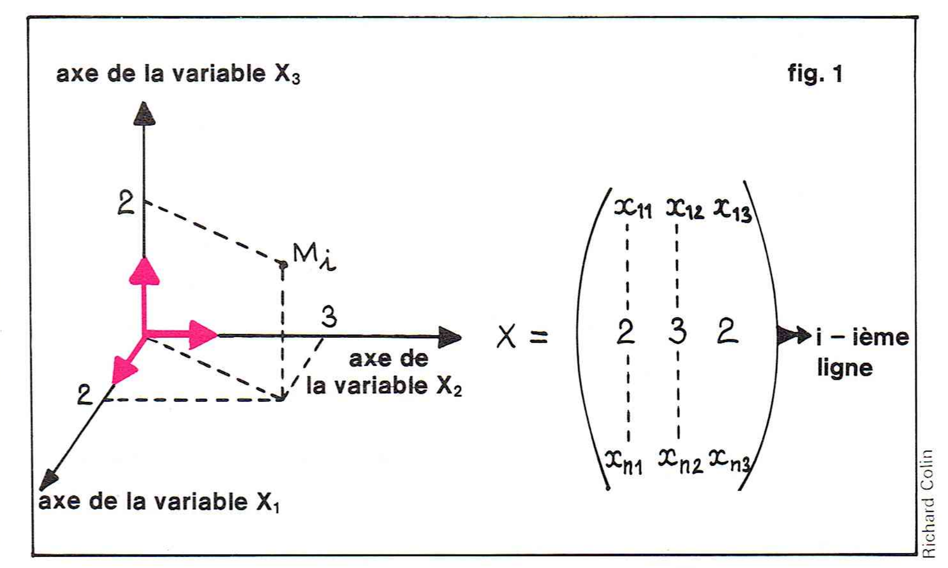
On choisit pour représenter les données l'espace qui a la dimension la plus petite et c'est en général celui des variables. Dans ce cas, on a alors un nuage de n « points individus » dans l'espace des variables $\mathbb{R}^p$ qu'on munit de la métrique euclidienne, et lorsque p est supérieur à 3, on ne peut visualiser les données.

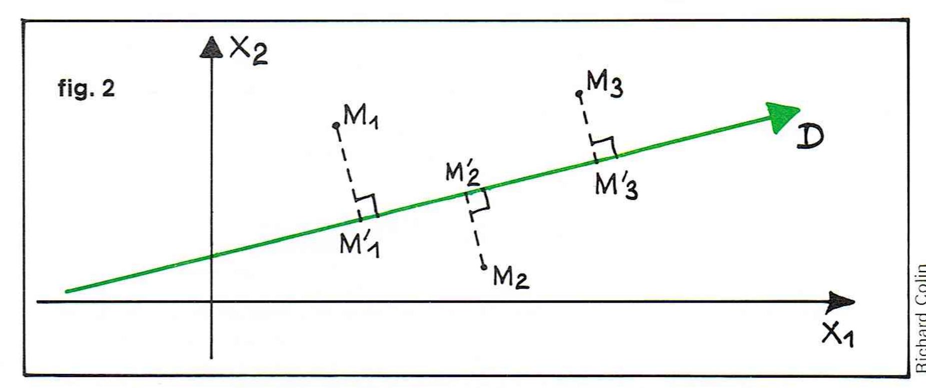
Le principe de l'analyse en composante principale consiste à remplacer les p variables par un petit nombre de nouvelles variables $F_1, F_2, ... F_k$ appelées *facteurs*, combinaisons linéaires des précédentes, sans perdre trop d'informations, ce qui revient géométriquement à projeter le nuage des n points sur un sous-espace de dimension aussi faible que possible sans trop de distorsion. Il s'agit donc de choisir un nouveau système d'axes (qu'on prendra orthogonaux) qui s'adapte le mieux possible à la forme du nuage.

Si on a, par exemple, trois individus et deux variables et si on suppose que les trois points M_1, M_2 et M_3 ne sont pas alignés dans l'espace des variables, le problème est de rechercher une droite D telle que les projections M_1', M_2' et M_3' donnent le plus de renseignements possibles sur les positions relatives de M_1, M_2 et M_3 (*fig.* 2).

On peut imaginer beaucoup de critères mathématiques rendant compte de la « fidélité » de la nouvelle représentation. Citons-en quelques-uns :

— la somme des longueurs M_iM_i' est minimale ;





▲ *En haut, figure 1 : un exemple de tableau* **X** *et sa représentation géométrique dans le cas de trois variables. Ci-dessus, figure 2 : recherche d'une droite* **D** *dans le cas de trois individus et de deux variables.*

— le maximum des longueurs M_iM_i' est minimal ;

— la somme des carrés des longueurs M_iM_i' est minimale ;

— si on accorde plus d'importance à certains individus plutôt qu'à d'autres, c'est-à-dire si on affecte à chaque individu un poids p_i, on peut minimiser :

$$\sum_i p_i M_iM_i' \quad \text{ou} \quad \sum_i p_i \overline{M_iM_i'}^2$$

Le *critère* choisi dans l'*analyse en composantes principales* est le suivant :

$$\text{maximiser} \sum_{i,k} \overline{M_i'M_k'}^2,$$

où M_i' et M_k' sont les projections orthogonales de M_i et M_k sur un sous-espace de $\mathbb{R}^p$ de faible dimension.

Appelons G le *centre de gravité* (ou *point moyen*) du nuage des points M_i (G a pour j-ième coordonnée $\left.\frac{1}{n}\sum_{i=1}^{n} x_{ij}\right)$. On montre facilement que :

$$\sum_{i,k} \overline{M_i'M_k'}^2 = 2\,n \sum_i \overline{G'M_i'}^2.$$

On en conclut que le critère de l'analyse en composantes principales est équivalent au critère suivant : la somme des carrés des projections des GM_i est maximale.

Recherche des axes factoriels

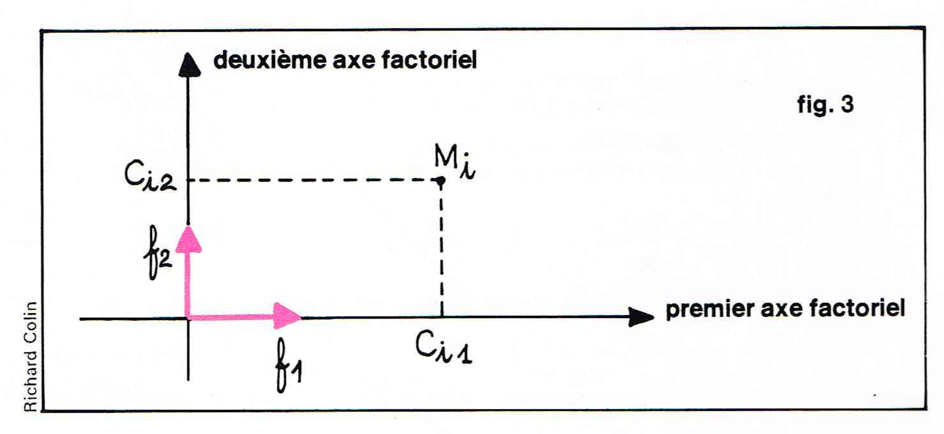
Appelons v_{jk} la *covariance* entre les variables X_j et X_k :

$$v_{jk} = \frac{1}{n}\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)\,(x_{ik} - \bar{x}_k)$$

et V la *matrice des covariances* des p variables :
V est une *matrice carrée* de dimension p, *symétrique*

$$(\text{covar}\ (X_j, X_k) = \text{covar}\ (X_k, X_j)),$$

positive (${}^t w\, V\, w \geqslant 0$ quel que soit le vecteur w de $\mathbb{R}^p$) et même définie, car si elle ne l'est pas, cela signifie qu'il



▲ ***Figure 3** : voir développement dans le texte* **Interprétation statistique.**

▼ ***Ci-dessous, figure 4** : dans ce cas, l'individu (2) est très bien représenté, tandis que (1) et (3) le sont moins bien.*
***En bas, figure 5** : dans cet exemple, la variable X_j, qui est proche du plan $(\vec{f_1}, \vec{f_2})$, est bien représentée, tandis que la variable X_k l'est moins bien.*

existe une même relation linéaire entre les variables pour chaque observation et qu'on a une variable expliquée par les $(p-1)$ autres; on ne considère alors que $(p-1)$ variables, ce qui entraîne qu'on n'envisagera que le cas où V est une *matrice symétrique définie positive*.

Une telle matrice possède p valeurs propres réelles, strictement positives. Classons les vecteurs propres unitaires $\vec{f_l}$ dans l'ordre décroissant des valeurs propres (si on a des valeurs propres multiples, le classement n'est pas unique; ce problème ne sera pas envisagé ici).

On démontre que : *le sous-espace à m dimensions (m < p) sur lequel le nuage se projette avec une déformation minimale est engendré par les m premiers vecteurs propres.*

On appelle *l-ième axe factoriel* l'axe défini par le l-ième vecteur propre, et le l-ième vecteur propre est appelé *l-ième composante principale*.

Interprétation statistique

Quel que soit le point M_i, il est possible de le repérer dans la base $\vec{f_1}, \vec{f_2}, ..., \vec{f_l}, ..., \vec{f_p}$. Appelons c_{il} la l-ième coordonnée de M_i dans cette nouvelle base (*fig.* 3).

Au l-ième axe factoriel, on peut donc associer une variable F_l qu'on appelle *l-ième facteur* et qui a pour composantes : $c_{1l}, c_{2l}, ..., c_{il}, ..., c_{nl}$.

On peut considérer les facteurs comme de nouvelles variables et l'ordre dans lequel sont rangés les c_{il} sur le l-ième axe factoriel peut permettre d'interpréter le l-ième facteur. Mais il faut être très prudent et s'assurer avant d'interpréter que les proximités constatées sur l'axe factoriel correspondent réellement à une proximité dans l'espace et non seulement à un « effet d'optique » dû à la projection. On montre que :

$$\cos(\overrightarrow{GM_i}, \vec{f_k}) = \frac{\overrightarrow{GM_i} \cdot \vec{f_k}}{\|\overrightarrow{GM_i}\|}$$

et si, par exemple, pour un point M_i, on a :

$$\cos^2(\overrightarrow{GM_i}, \vec{f_1}) + \cos^2(\overrightarrow{GM_i}, \vec{f_2}) \# 1$$

on pourra en conclure que M_i est très proche du plan des deux premiers axes factoriels. Sur la *figure* 4, l'individu **(2)** est parfaitement représenté, tandis que **(1)** et surtout **(3)** le sont moins bien.

On démontre que la *moyenne* d'un facteur est égale à la composante de la projection du point moyen G sur l'axe factoriel qui lui correspond et que sa *variance* est égale à la valeur propre associée. On a ainsi sur le premier axe factoriel la plus grande dispersion des projections des « points individus ». On peut donc dire que les axes factoriels ont la propriété d'extraire progressivement le plus d'informations possibles sur l'ensemble du nuage. Le coefficient choisi pour mesurer la fidélité de la projection du nuage sur les m premiers axes factoriels est :

$$\frac{\sum_{l=1}^{m} \lambda_l}{\sum_{l=1}^{p} \lambda_l} = \frac{\sum_{l=1}^{m} \lambda_l}{\text{trace de V}}$$

coefficient égal à 1 pour $m = p$ (déformation nulle).

On peut représenter les variables sur un axe factoriel ou un plan factoriel en projetant sur cet axe ou ce plan leurs axes et vecteurs unitaires (*fig.* 5).

Soit e'_j la projection du vecteur unitaire de la j-ième variable. On montre facilement que sa composante sur le l-ième axe factoriel est f_{lj}. Si la longueur de e'_j est voisine de 1 dans le plan des deux premiers axes factoriels, cela signifie que l'axe de la variable X_j est proche du plan de ces deux premiers axes. Il est donc intéressant de considérer les positions des extrémités des vecteurs e'_j par rapport au cercle de centre 0 et de rayon 1. Sur la *figure* 5, la variable X_j est bien représentée : elle est proche du plan $(\vec{f_1}, \vec{f_2})$, tandis que la variable X_k l'est moins bien.

On démontre que le coefficient de corrélation entre la variable X_j et le facteur F_l est égal à :

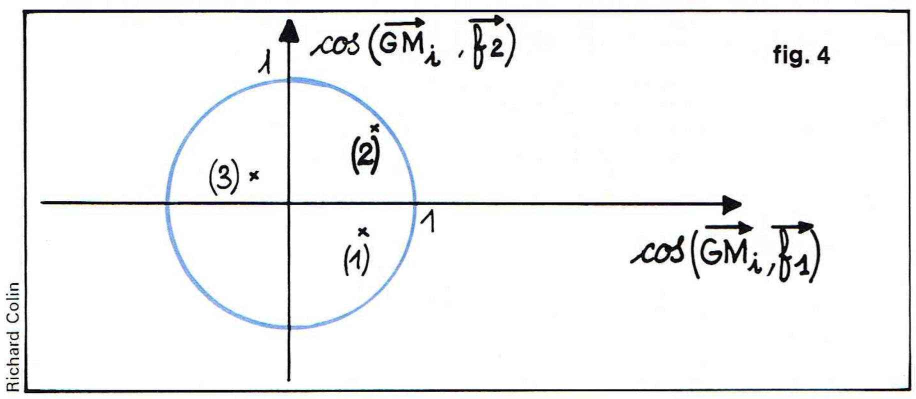
$$r(X_j, F_l) = \frac{\sqrt{\lambda_l}}{\sqrt{v_{jj}}} f_{lj}$$

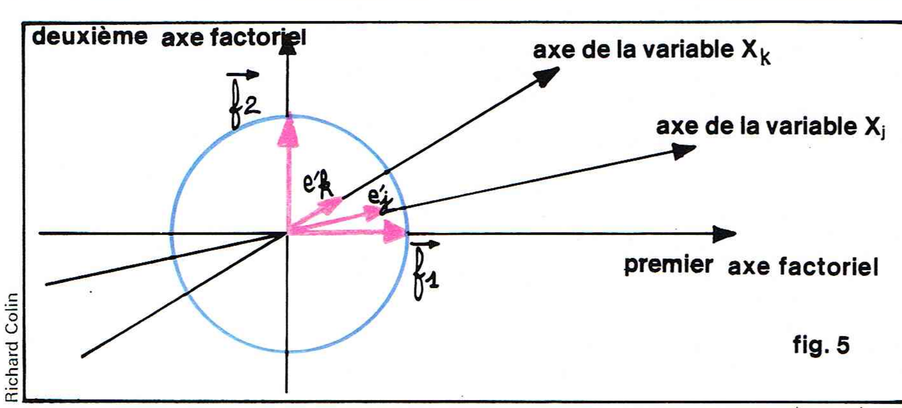
Si toutes les variables ont des dispersions voisines, on pourra donc dire que le l-ième facteur est le plus corrélé avec les variables pour lesquelles les f_{lj} sont les plus grands en valeur absolue (corrélation directe si les f_{lj} sont positifs et inverse s'ils sont négatifs).

Si on a des points M'_i voisins de l'axe de la variable X_j dans le plan des deux premiers axes factoriels, cela peut vouloir signifier que M_i est proche de l'axe de la j-ième variable dans $\mathbb{R}^p$ si le nuage n'est pas trop déformé. Ceci implique alors que la variable X_j caractérise le i-ième individu, mais il vaut mieux prendre la précaution de vérifier que cette proximité a toujours lieu dans le plan des 3e et 4e axes factoriels, car un point M_i peut être situé en projection sur l'axe de la variable X_j, tout en étant fort éloigné de cet axe dans l'espace $\mathbb{R}^p$.

Un axe factoriel peut donc être interprété à l'aide des variables ou par l'intermédiaire de l'échantillon d'individus envisagé, cette dernière interprétation ayant l'avantage d'être moins tautologique que celle qui s'appuie sur les variables elles-mêmes; cependant il faut avoir conscience que le lien établi entre deux familles de variables à l'aide d'un échantillon d'individus risque de n'être qu'accidentel et il est nécessaire d'être prudent avant de tirer une conclusion.

Lorsque les données sont hétérogènes, l'analyse en composantes principales est plus intéressante si elle est faite sur des variables centrées réduites, ceci afin d'éliminer les effets de taille.





Analyse factorielle des correspondances

Nous allons la présenter à partir d'un exemple. Supposons que l'on s'intéresse aux résultats d'élections dans les diverses circonscriptions de Paris; on dresse alors un tableau $K = (k_{ij})$ où l'on trouve à l'intersection de la ligne i et de la colonne j le nombre de suffrages obtenus par le candidat j dans la circonscription i.

L'analyse des correspondances permet de comparer les circonscriptions (deux circonscriptions se ressembleront d'autant plus que les pourcentages de chaque candidat par rapport au nombre de suffrages comptabilisés dans chacune d'elles seront plus proches) ou les candidats (deux candidats sont d'autant plus proches que les pourcentages des suffrages obtenus dans chaque circonscription par rapport au nombre total des suffrages obtenus par chacun d'eux sont plus voisins).

Deux *remarques* s'imposent ici :

— l'ensemble I des n circonscriptions et l'ensemble J des p candidats jouent des rôles tout à fait symétriques;

— la comparaison des lignes ou des colonnes se fait non sur les données brutes mais sur les profils de ces lignes ou de ces colonnes.

Si l'on pose : $k = \sum_{i,j} k_{ij}$, $f_{ij} = \frac{k_{ij}}{k}$ $\quad f_i = \sum_j f_{ij}$

le profil de la circonscription i est le vecteur

$$f^i_J = \frac{1}{f_i} f_{iJ} \quad \text{avec} \quad f_{iJ} = (f_{i1}, f_{i2}, ..., f_{ip})$$

On considère le nuage des n circonscriptions :

$$\mathcal{N}(I) = \{(f^i_J, f_i) \mid i = 1, ..., n\}$$

où f^i_J est le vecteur déjà cité et f_i un coefficient de pondération et symétriquement le nuage :

$$\mathcal{N}(J) = \{(f^j_I, f_j) \mid j = 1, ..., p\}$$

$$\text{où } f^j_I = \frac{1}{f_j}\ {}^t(f_{1j}, f_{2j}, ..., f_{nj}) \qquad f_j = \sum_i f_{ij}.$$

La distance du χ^2 définie par

$$d^2(i,i') = \sum_j \frac{1}{f_j}\left(\frac{f_{ij}}{f_i} - \frac{f_{i'j}}{f_{i'}}\right)^2 = \left\| f^i_J - f^{i'}_J \right\|^2_{f_J}$$

$$d^2(j,j') = \sum_i \frac{1}{f_i}\left(\frac{f_{ij}}{f_j} - \frac{f_{ij'}}{f_{j'}}\right)^2 = \left\| f^j_I - f^{j'}_I \right\|^2_{f_I}$$

traduit la proximité que nous avons décrite dans la présentation et surtout qui respecte le **principe d'équivalence distributionnelle,** à savoir : *si deux vecteurs colonnes (ou lignes) ont même profil et si on leur substitue dans le tableau une colonne (ou une ligne) unique égale à leur somme, alors la distance distributionnelle entre éléments de I (de J) n'est pas modifiée.*

Cette propriété est fondamentale car elle rend la méthode peu sensible aux partitions adoptées dans les ensembles I et J. Dans l'exemple que nous avons choisi, le découpage des ensembles I et J s'imposait mais ce n'est pas toujours le cas. Ainsi, imaginons une enquête sur la consommation alimentaire des ménages; faut-il mettre sous des rubriques différentes les dépenses concernant les boissons gazeuses et les jus de fruits? C'est inutile, si ces dépenses ont des profils voisins (c'est-à-dire représentent des comportements similaires); cela peut être intéressant sinon, mais la question ne se pose plus si l'on a choisi comme distance celle du χ^2. La stabilité des résultats obtenus en analyse des correspondances provient de ce qu'ils sont relativement indépendants de l'arbitraire des nomenclatures.

Si on modifie l'échelle des axes, plus précisément si on représente la i-ième ligne par le vecteur

$$OM_i = \left(\frac{f_{i1}}{f_i\sqrt{f_1}}, \frac{f_{i2}}{f_i\sqrt{f_2}}, ..., \frac{f_{ip}}{f_i\sqrt{f_p}}\right) = (x_{i1}, ..., x_{ip})$$

alors $d^2(i,i') = \sum_j (x_{ij} - x_{i'j})^2$

comme dans l'analyse en composantes principales.

L'analyse des correspondances revient à effectuer une analyse en composantes principales sur des points dont les coordonnées ont été modifiées de telle sorte que la distance entre ces points, calculée sur le tableau initial, soit une distance du χ^2.

Analyser la correspondance entre I et J, c'est mesurer à quel point le phénomène étudié diffère de l'indépendance statistique entre I et J.

S'il y avait indépendance statistique, c'est-à-dire si pour tout i de l'ensemble I et tout j de l'ensemble J $f_{ij} = f_i f_j$, alors :

$$OM_i = \left(\frac{f_i f_1}{f_i\sqrt{f_1}}, \frac{f_i f_2}{f_i\sqrt{f_2}}, ..., \frac{f_i f_p}{f_i\sqrt{f_p}}\right) = \left(\sqrt{f_1}, \sqrt{f_2}, ..., \sqrt{f_p}\right) = OG$$

tous les points du nuage seraient confondus avec le centre de gravité $G = \sum_i f_i M^i$.

Ceci est généralement faux, et étudier la dépendance entre I et J revient à déterminer la « forme » du nuage $\mathcal{N}$ (I) ou $\mathcal{N}$ (J) : recherche des directions d'allongement, axes principaux d'inertie, etc.

C'est encore, dans le cas où le tableau analysé est un tableau de contingence, mesurer en quoi la quantité :

$$d^2 = \sum_{i,j} k \frac{(f_{ij} - f_i f_j)^2}{f_i f_j}$$

réalisation d'un $\chi^2_{(p-1)\,(n-1)}$, diffère significativement de zéro.

On démontre que l'inertie totale du nuage $\mathcal{N}$ (I), de même que celle du nuage $\mathcal{N}$ (J), est égale à $\frac{d^2}{k}$, la matrice d'inertie H étant $R\,{}^tR$ où R est une matrice de terme général

$$r_{ij} = \frac{f_{ij} - f_i f_j}{\sqrt{f_i f_j}},$$

la matrice d'inertie du nuage $\mathcal{N}$ (J) est ${}^tR\, R = S$.

Les valeurs propres non nulles de H et S sont les mêmes; de plus, si u est un vecteur propre de H, on a

$$Hu = R\,{}^tRu = \lambda u$$

d'où, en multipliant par ${}^tR$:

$$({}^tR\, R)\,{}^tRu = \lambda\,{}^tRu$$

${}^tRu$ est vecteur propre de S.

Il en résulte une relation linéaire particulièrement simple entre les facteurs des deux analyses.

La symétrie des opérations effectuées dans l'un ou l'autre nuage a pour conséquence que l'analyse en composantes principales peut se faire indifféremment dans l'espace des individus ou dans celui des variables. Pratiquement, cela permet de choisir l'espace de dimension la plus faible.

On peut signaler rapidement quelques *résultats* complémentaires : le nuage $\mathcal{N}$ (I) appartient à l'hyperplan orthogonal à OG. OG est vecteur propre de H, associé à la valeur propre 0; la recherche des vecteurs propres et valeurs propres de H se ramène à celle des vecteurs propres et valeurs propres de W où :

$$W_{kl} = \sum_i \frac{f_{ik}\, f_{il}}{f_i\sqrt{f_l}\sqrt{f_k}}$$

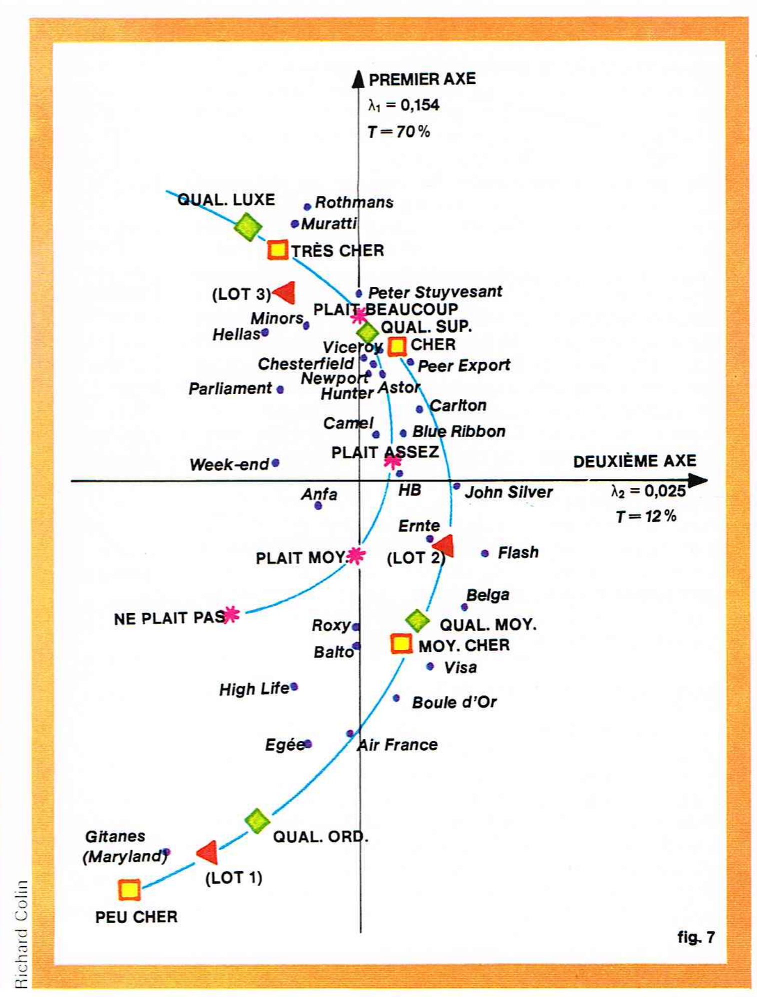
OG étant le vecteur propre de W associé à la valeur propre 1. Toutes les autres valeurs propres de W sont comprises entre 0 et 1.

La représentation simultanée des deux nuages sur les plans factoriels, la possibilité de projeter sur ces plans des individus (ou des variables) supplémentaires améliorent la puissance de la méthode.

L'interprétation des facteurs se fait à partir des graphiques (plans des axes 1 — 2, 1 — 3, 2 — 3, etc.) à l'aide des listes de contributions qui mesurent l'importance de chaque élément dans l'apparition d'un facteur, ou rendent compte de la contribution du facteur à tel ou tel élément.

L'exemple choisi (comparaison des résultats d'élection dans différentes circonscriptions) est celui d'un tableau de contingence, mais l'analyse des correspondances a de multiples applications : tableaux de mesure, tableaux de note d'intensité, tableaux de description logique (tableaux remplis de 1 et de 0 suivant la présence ou l'absence d'un certain caractère pour un individu donné). On démontre dans ce dernier cas que si le tableau logique est mis sous forme disjonctive complète, son analyse est équivalente à celle du tableau de contingence associé.

fig. 6 MARQUE \ ATTRIBUT	Qualité				Agréable				Cher				Lot		
	ord.	moy.	sup.	lux.	non	moy.	assez	très	peu	moy.	assez	très	1	2	3
Air France	15	65	13	7	20	50	28	2	22	55	22	1	29	61	10
Anfa	12	48	25	15	27	26	27	20	7	40	32	21	18	47	35
Astor	7	30	45	18	7	33	46	14	3	23	55	19	0	37	63
Balto	14	57	25	4	19	35	34	12	20	48	29	3	16	52	32
Belga	16	64	10	10	7	41	43	9	7	53	37	3	14	63	18
Blue Ribbon	6	39	37	18	7	34	35	24	4	35	51	10	5	58	37
Boule d'or	33	49	15	3	13	52	27	8	15	62	22	1	37	48	15
Camel	12	46	28	14	11	37	32	20	3	27	46	24	3	25	72
Carlton	5	37	45	13	10	29	42	19	1	31	55	13	14	52	34
Chesterfield	4	36	34	26	10	32	28	30	2	17	59	22	0	18	82
Egée	40	51	7	2	18	36	35	11	28	52	19	1	45	40	15
Ernte 23	11	60	25	4	9	39	40	12	8	33	47	12	7	68	25
Flash	12	62	16	10	5	41	42	12	3	46	47	3	9	69	22
Gitanes Maryland	44	50	5	1	22	52	23	3	46	45	7	2	56	32	12
H. B.	13	49	15	13	14	34	28	24	4	31	54	11	16	59	25
Hellas	2	27	27	44	12	37	32	19	3	22	46	29	10	50	40
High Life	16	63	15	6	14	43	30	13	32	36	28	4	29	40	36
Hunter	7	34	34	25	5	32	39	24	5	21	62	12	16	46	38
John Silver	6	57	30	7	3	32	39	26	5	33	53	9	6	51	43
Minors	3	19	45	33	11	40	30	19	3	20	56	21	3	26	71
Muratti	1	10	45	44	5	23	37	35	1	12	65	22	2	17	81
Newport	6	33	33	28	16	29	33	22	0	21	67	12	6	51	43
Parliament	12	26	35	27	14	29	43	14	10	21	48	21	11	40	49
Peer Export	3	24	58	15	4	26	41	29	3	32	56	9	2	43	55
Peter Stuyvesant	2	24	49	25	5	33	32	30	1	18	61	20	4	31	65
Rothmans	1	11	43	45	4	20	41	35	0	10	69	21	1	25	74
Roxy	17	55	21	7	14	37	37	12	21	42	33	4	16	58	26
Viceroy	2	35	42	21	5	28	43	24	2	23	52	23	2	44	54
Visa	12	63	27	8	10	46	38	6	13	69	17	1	24	60	16
Week-End	6	29	38	27	21	29	33	17	11	42	30	17	15	31	54



Richard Colin

▲ *Figures 6 et 7 : deux illustrations tirées de l'ouvrage de J.-P. Benzecri et coll. : l'*Analyse des correspondances*, tome II de l'*Analyse des données *(Dunod) et portant sur l'analyse de 30 marques de cigarettes.*

L'analyse des correspondances issue de l'étude de tableaux de contingence suppose que le tableau initial soit un *tableau de nombres positifs*. Si tel n'est pas le cas, on peut procéder par dédoublement des colonnes.

L'analyse des correspondances est due au professeur J.-P. Benzecri (voir *Bibliographie*).

Exemple d'application : analyse de 30 marques de cigarettes (*fig.* 6 et 7). Chaque marque a été testée auprès de 100 fumeurs interrogés sur ses caractéristiques de qualité, d'agrément et de coût. La cohérence des réponses a été vérifiée par une comparaison avec 3 lots de paquets de cigarettes de qualité différente.

L'analyse des correspondances donne de très bons résultats : en effet les deux premiers axes totalisent 82 % de l'inertie du nuage.

Sur le schéma on remarque que le premier axe classe dans l'ordre croissant les différentes modalités de la qualité, de l'agrément et du prix, et que le deuxième axe oppose les valeurs extrêmes (peu cher, très cher, ordinaire, luxe) aux valeurs moyennes.

Analyse discriminante

On suppose que sur une population répartie *a priori* en q classes disjointes, on a mesuré p variables. L'objectif de l'analyse discriminante est de rechercher sur les variables une formule, linéaire ou autre, qui permette de déterminer la classe à laquelle doit être affecté chaque individu.

Par exemple, le dossier médical d'un sujet i (n sujets) contient les résultats numériques de p examens

$$(x_{i1}, x_{i2}, \ldots, x_{ip})$$

et un diagnostic y_i (q modalités). Une analyse discriminante indiquera les examens qui permettent le mieux de différencier les diagnostics et donc éventuellement de déceler ceux qui sont le moins utiles ; d'autre part la formule établie sur les variables les plus discriminantes donnera le diagnostic de tout nouveau malade.

Les n individus peuvent être représentés dans $\mathbb{R}^p$, muni de la métrique usuelle (*fig.* 8). A chaque point x_i on affectera la masse $m_i = \frac{1}{n}$. Le problème revient à chercher dans $\mathbb{R}^p$ la direction qui sépare au mieux les q groupes.

Si on désigne par I l'ensemble des individus, pour u fixé,

$V_I(u) = \sum_{i \in I} m_i \, u(x_i)^2$ est la **variance totale,**

$V_{g_r}(u) = \sum_{x_i \in g_r} m_i \, u(x_i - \bar{g}_r)^2$ est la **variance intra-classe** de la classe g_r,

$V_{\mathcal{G}}(u) = \sum_{g_r \in \mathcal{G}} m_{g_r} \, u(\bar{g}r - \bar{g})^2$ est la **variance inter-classes,**

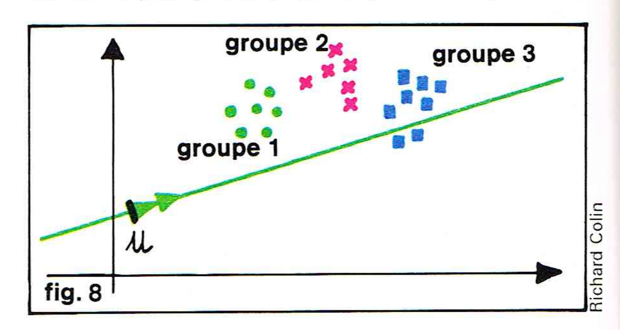
avec $u(t) = \sum_{j=1}^{p} u_j t_j$

$\bar{g}_r$ centre de gravité de la classe g_r

$m_{g_r} = \sum_{x_i \in g_r} m_i$ masse de la classe g_r

$\bar{g}$ centre de gravité de $\mathcal{G} = \{g_1, g_2, \ldots, g_q\}$

Le critère choisi consiste à rendre la variance inter-classes la plus grande possible (centres de gravité des



Richard Colin

▶ *Figure 8 : la direction* $\vec{u}$ *sépare les trois groupes.*

classes espacés) et les variances inter-classes les plus faibles possibles (classes peu dispersées). Plus précisément on cherchera à maximiser le rapport

$$\frac{V_{\mathcal{G}}(u)}{\sum_{g_r \in \mathcal{G}} V_{g_r}(u)}$$

Or on sait d'après le théorème de Huyghens que :

$$V_I(u) = V_{\mathcal{G}}(u) + \sum_{g_r \in \mathcal{G}} V_{g_r}(u)$$

Le problème posé équivaut à maximiser $\frac{V_{\mathcal{G}}(u)}{V_I(u)}$, u étant un vecteur unitaire.

Si on désigne par T la matrice des variances totales et par B la matrice des variances inter-classes, la solution du problème sera telle que $T^{-1}Bu = \lambda u$, λ étant la plus grande valeur propre associée à la matrice $T^{-1}B$.

On est ramené au problème suivant : effectuer une analyse en composantes principales sur le nuage $\mathcal{G}$ des centres de gravité de forme quadratique d'inertie B, l'espace vectoriel des individus I étant muni de la métrique T^{-1}.

Dans la pratique, n (nombre des individus étudiés) doit surpasser p (nombre de paramètres de dispersion), sinon on ne connaît pas la dispersion des classes autour de leur centre et on ne peut savoir à quelle classe rattacher un nouvel individu. Diverses méthodes de calcul ont été proposées (Fischer, Mahalanobis).

Régression linéaire

La régression linéaire répond à un problème moins descriptif que l'analyse factorielle : on a n observations sur $p+1$ variables (en général n est nettement plus grand que p), et il s'agit d'examiner comment on peut expliquer une de ces variables [la $(p+1)$-ième par exemple] à l'aide d'une combinaison linéaire des p autres.

La représentation géométrique la plus simple consiste à se placer dans l'espace euclidien orthonormé $\mathbb{R}^n$ en considérant chacune des variables x comme un point de $\mathbb{R}^n$ dont les composantes dans la base de $\mathbb{R}^n$ sont les valeurs que cette variable a prises respectivement pour les n observations. Pour la commodité des développements ultérieurs on supposera que toutes les *variables* sont *centrées*. Mentionnons que l'opération de centrage s'interprète géométriquement comme la projection orthogonale des variables non centrées sur l'hyperplan orthogonal à la « bissectrice » des axes (droite de vecteur directeur $\left\{\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, ..., \frac{1}{\sqrt{n}}\right\}$ dans la base choisie).

Ces variables $x_1, x_2, ..., x_p$ engendrent un sous-espace vectoriel W de $\mathbb{R}^n$ dans lequel il s'agit de trouver le point $\hat{y}$ « le plus proche possible » du point y représentant la $(p+1)$-ième variable (*fig.* 9).

L'intention explicative, voire prédictive, d'un tel modèle est claire : connaître la valeur $\hat{y}_i$ que $\hat{y}$ a prise à la i-ième observation permet d'approcher la valeur y_i de y pour cette même observation. Mais la précision de cette approximation dépend évidemment de la distance de y à $\hat{y}$ et de i. La variable y peut aussi se prêter plus ou moins bien à une régression linéaire en l'ensemble des variables x_j. Inversement, le fait de trouver une bonne régression ne doit pas entraîner imprudemment à des conclusions sur les variables considérées.

On suppose que $y_1, y_2, ..., y_i, ..., y_n$ sont les réalisations d'un échantillon de taille n d'une variable aléatoire Y, de même que $x_{j1}, x_{j2}, ..., x_{ji}, ..., x_{jn}$ celles d'une variable aléatoire X_j. Le modèle linéaire suppose qu'entre les p variables explicatives $X_1, ..., X_j, ..., X_p$ et la variable expliquée Y existe la relation suivante :

$$E(Y/X_1, X_2, ..., X_j, ..., X_p) = \sum_{j=1}^{p} \beta_j X_j,$$

donc que :

$$Y = \sum_{j=1}^{p} \beta_j X_j + \varepsilon$$

Cela revient à supposer que Y ne diffère d'une combinaison linéaire des X_j que par un résidu aléatoire d'espérance nulle.

Puisqu'on dispose d'un échantillon de taille n de chacune des variables Y, $X_1, ..., X_j, ..., X_p$, on a les relations suivantes :

$$\begin{aligned} y_1 &= \beta_1 x_{11} + \beta_2 x_{21} + ... + \beta_j x_{j1} + ... + \beta_p x_{p1} + e_1 \\ &\vdots \\ y_i &= \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_j x_{ji} + ... + \beta_p x_{pi} + e_i \\ &\vdots \\ y_n &= \beta_1 x_{1n} + \beta_2 x_{2n} + ... + \beta_j x_{jn} + ... + \beta_p x_{pn} + e_n \end{aligned}$$

Ce système d'équation peut s'écrire sous forme matricielle :

$$\vec{y} = X\vec{\beta} + \vec{e}$$

où :

$$\begin{aligned} {}^t\vec{y} &= (y_1, y_2, ..., y_i, ..., y_n), \\ {}^t\vec{\beta} &= (\beta_1, \beta_2, ..., \beta_j, ..., \beta_p), \\ {}^t\vec{e} &= (e_1, e_2, ..., e_i, ..., e_n), \end{aligned}$$

et

$$X = \begin{pmatrix} x_{11} & ... & x_{j1} & ... & x_{p1} \\ x_{1i} & ... & x_{ji} & ... & x_{pi} \\ x_{1n} & ... & x_{jn} & ... & x_{pn} \end{pmatrix}$$

Pour estimer β, on est obligé de faire des hypothèses sur les résidus et les variables explicatives. Si on peut sans trop de risque faire des hypothèses assez fortes sur les résidus, on emploie la *méthode des « moindres carrés ordinaires » (M.C.O.)* qui donne pour estimation de $\vec{\beta}$:

$$\vec{b} = ({}^tXX)^{-1}\, {}^tX\vec{y}$$

Cette méthode revient à calculer $\vec{b}$ qui minimise :

$$\|\vec{y} - X\vec{b}\|^2 = {}^t(\vec{y} - X\vec{b})\,(\vec{y} - X\vec{b})$$

qui n'est autre que le carré de la distance euclidienne de y à la variété linéaire engendrée par les p vecteurs colonnes de X.

Sous ces mêmes hypothèses, on démontre que l'estimateur associé à l'estimation $\vec{b}$ est un « bon estimateur », c'est-à-dire qu'il est sans biais et convergent.

Lorsque les hypothèses sur les résidus ne sont manifestement pas remplies, on emploie une méthode qui en demande de moins fortes, mais qui est plus compliquée : *« la méthode des moindres carrés généralisée » (M.C.G.)*.

Remarque : dans certains problèmes de régression, il peut être nécessaire d'imposer des contraintes : par exemple, si les coefficients β_j représentent des prix, ils doivent être positifs ou nuls, s'ils représentent des probabilités, ils doivent être compris entre 0 et 1 ; on est alors dans le domaine de la *programmation quadratique*.

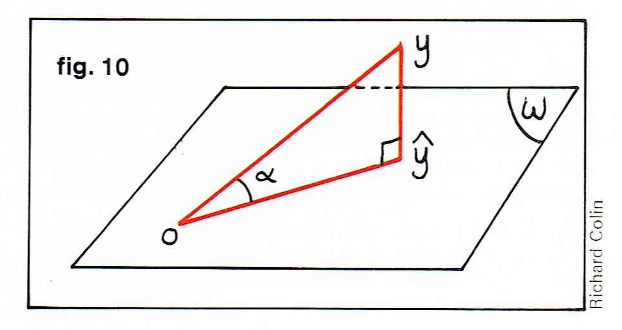
Il est naturel de dire que la régression est d'autant « meilleure » que $\|e\| = \|y - \hat{y}\|$ est plus faible.

On mesure la qualité de la régression par le rapport :

$$R = \frac{\|\hat{y}\|}{\|y\|};$$

on a $0 \leqslant R \leqslant 1$ et la régression est évidemment d'autant

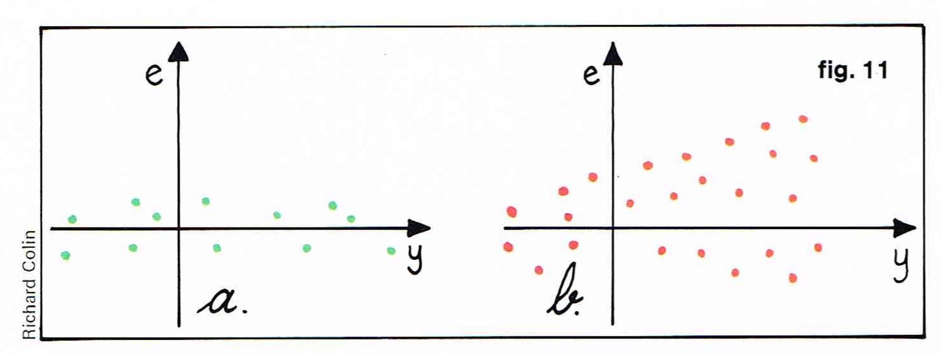
fig. 9

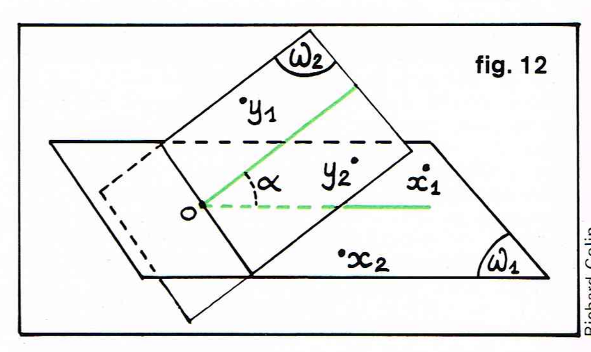


◄ *A gauche, figure 9 : les variables $x_1, x_2, ..., x_p$ engendrent un sous-espace vectoriel W de $\mathbb{R}^n$ dans lequel il s'agit de trouver le point $\hat{y}$ « le plus proche possible » du point y représentant la $(p+1)$-ième variable. A droite, figure 10 : $R = \left\|\frac{\hat{y}}{y}\right\|$ est donc analogue à un coefficient de corrélation puisque si $\mathbb{R}^n$ est muni de la métrique euclidienne, R = cos α.*

Richard Colin

Richard Colin





▲ *A gauche, figure 11 : a, tous les y sont approximés avec la même précision, qui est bonne ; b, l'approximation est d'autant plus mauvaise que y_i est grand — il faut donc modifier le modèle. A droite, figure 12 : représentation géométrique dans l'espace $\mathbb{R}^n$ des individus.*

meilleure que R est voisin de 1. Puisque $\mathbb{R}^n$ est muni de la métrique euclidienne, R n'est autre que le cosinus de l'angle $\widehat{yOy}$. R est donc analogue à un coefficient de corrélation (*fig.* 10).

Si les variables ont été préalablement centrées :

$$R = \frac{\|\vec{\hat{y}}\|}{\|\vec{y}\|} = \frac{\|\vec{\hat{y}}\|^2}{\|\vec{y}\|\,\|\vec{\hat{y}}\|} = \frac{\text{covar}\,(y, \hat{y})}{\sigma_y \cdot \sigma_{\hat{y}}}$$

car :

$$\|\vec{y}\| = \sqrt{\sum_{i=1}^{n} y_i^2} = \sqrt{n}\,\sigma_y ;$$

$$\|\vec{\hat{y}}\| = \sqrt{\sum_{i=1}^{n} \hat{y}_i^2} = \sqrt{n}\,\sigma_{\hat{y}} ;$$

et $\vec{y} - \vec{\hat{y}}$ étant orthogonal à $\vec{\hat{y}}$, on a :

$\|\vec{\hat{y}}\|^2 = \|\vec{y}\|^2 - \|\vec{y} - \vec{\hat{y}}\|^2$, ce qui entraîne que :

$$\|\vec{\hat{y}}\|^2 = \sum_{i=1}^{n} y_i\hat{y}_i = n \text{ covar } (y, \hat{y}).$$

On déduit de ces résultats que R n'est autre que le coefficient de corrélation entre y et $\hat{y}$. On l'appelle *coefficient de corrélation multiple* entre la variable y et l'ensemble $\{x_1, x_2, ..., x_j, ..., x_p\}$ des variables explicatives.

On utilise aussi un autre rapport pour mesurer la qualité de la régression :

$$F = \frac{\|\vec{\hat{y}}\|^2/p}{\|\vec{y} - \vec{\hat{y}}\|^2/n-p} = \frac{n-p}{p}\,\frac{\|\vec{\hat{y}}\|^2}{\|\vec{y}\|^2 - \|\vec{\hat{y}}\|^2} = \frac{n-p}{p}\,\frac{R^2}{1-R^2}$$

Considéré comme une statistique d'échantillon, ce rapport suit, moyennant certaines hypothèses, une *distribution de Fischer-Snédecor* à p et $(n-p)$ degrés de liberté et, par référence à une table donnant la distribution de F, on peut juger de la « signification » de la régression. Ce rapport permet aussi de juger de l'intérêt d'inclure certaines variables dans la régression en mesurant « le pouvoir explicatif » de $(p-q)$ variables restantes alors que les q premières ont été utilisées.

Les rapports R et F ne donnent pas un examen détaillé de la qualité de la régression. Pour se rendre compte de l'approximation faite quand on remplace y_i par $\hat{y}_i$, il faut examiner la variation des quantités $e_i = y_i - \hat{y}_i$ en fonction de i. Pour cela, la méthode est de représenter dans un plan les couples (y_i, e_i), ce qui permet de se rendre compte si le résidu e_i est influencé par la valeur de y_i (*fig.* 11).

Il y a d'autres méthodes statistiques qui s'apparentent à la régression linéaire : la **régression polynômiale** qui se résout avec les mêmes techniques que la régression linéaire et l'**analyse de la variance** qui utilise des modèles de régression avec des variables explicatives qualitatives.

Il existe aussi des méthodes de **régression multiple** utilisées lorsqu'on a plusieurs variables à expliquer à l'aide des mêmes variables explicatives.

Analyse canonique

On peut considérer que l'analyse canonique généralise la régression linéaire, en ce sens qu'elle consiste à confronter deux groupes de variables quantitatives qui ont été mesurées sur n individus :

— p variables $X_1, X_2, ..., X_p$ appartenant à un certain domaine (par exemple économique),

— q variables $Y_1, Y_2, ..., Y_q$ appartenant à un autre domaine (par exemple sociologique) ;

et à trouver jusqu'à quel point on peut prévoir l'un des deux groupes à partir de l'autre.

On place les données dans un tableau R à n lignes et $(p+q)$ colonnes après avoir centré les variables :

$$R = \begin{pmatrix} x_{11} & x_{21} & \dots & x_{p1} & y_{11} & y_{21} & \dots & y_{q1} \\ x_{1i} & x_{2i} & \dots & x_{pi} & y_{1i} & y_{2i} & \dots & y_{qi} \\ \vdots & & & & \vdots & & & \\ x_{1n} & x_{2n} & \dots & x_{pn} & y_{1n} & y_{2n} & \dots & y_{qn} \end{pmatrix}$$

On adopte la représentation géométrique dans l'espace $\mathbb{R}^n$ des individus comme pour la régression linéaire (*fig.* 12).

En généralisant l'idée développée dans le modèle de la régression, il s'agit de trouver un vecteur $\vec{U_1}$ du sous-espace W_1 engendré par les variables x et un vecteur $\vec{U_2}$ du sous-espace W_2 engendré par les variables y de telle sorte que $\vec{U_1}$ et $\vec{U_2}$ fassent entre eux l'angle le plus petit possible.

Statistiquement, cela revient à chercher deux variables, chacune étant une combinaison linéaire d'un des deux groupes de variables considérées, qui soient le plus correlées possible. Ceci peut permettre, mais avec précaution, de prévoir comment se comporte, par exemple, un individu sur U_2 si on connaît son comportement selon « l'aspect » U_1.

Notons déjà que si W_1 et W_2 sont identiques, alors toute combinaison linéaire des x est combinaison linéaire des y et réciproquement. Par contre, si $W_1 \cap W_2 = \varnothing$, toute combinaison linéaire des y est orthogonale à toute combinaison linéaire des x et on dit alors que les variables y et x sont totalement différentes.

On vient de considérer deux situations extrêmes entre lesquelles il existe une infinité de situations intermédiaires possibles.

On démontre que la solution du problème posé s'obtient par le calcul des valeurs propres et vecteurs propres des matrices $V_{xx}^{-1}\,V_{xy}\,V_{yy}^{-1}\,V_{yx}$ et $V_{yy}^{-1}\,V_{yx}\,V_{xx}^{-1}\,V_{xy}$ où V_{xx} et V_{yy} désignent respectivement les matrices des covariances des variables x et des variables y, et V_{xy} désigne une matrice à p lignes et q colonnes dont le terme général de la i-ième ligne et j-ième colonne est égal à covar (x_i, y_j) [V_{yx} se définit d'une façon semblable]. Ces deux matrices ne sont pas du même ordre si $p \neq q$, mais sont de même rang. Elles correspondent aux opérateurs produit des projecteurs de W_1 sur W_2 et de W_2 sur W_1.

On démontre que ces deux matrices ont des valeurs propres égales toutes positives ou nulles et que le nombre de leurs valeurs propres égales à 1 est égal à la dimension de $W_1 \cap W_2$.

On démontre aussi que le premier couple $(\vec{U_1}, \vec{U_2})$, solution du problème, correspond respectivement aux vecteurs propres unitaires de

$$V_{xx}^{-1}\,V_{xy}\,V_{yy}^{-1}\,V_{yx} \text{ et } V_{yy}^{-1}\,V_{yx}\,V_{xx}^{-1}\,V_{xy}$$

associés à la plus grande valeur propre différente de 1. On l'appelle premier couple de *variables canoniques*, et la valeur propre correspondante est égale au carré du coefficient de corrélation linéaire entre ces deux variables.

On peut évidemment, comme en analyse factorielle, poursuivre l'étude en cherchant les couples suivants de variables associées $(U_1^{(2)}, U_2^{(2)})$, $(U_1^{(3)}, U_2^{(3)})$, ... On

démontre que $U_1^{(k)}$ et $U_1^{(l)}$ associés à des valeurs propres distinctes sont non correlés, de même que $U_1^{(k)}$ et $U_2^{(l)}$.

Pour interpréter les variables canoniques et par exemple U_1 on peut calculer les corrélations de chacune des variables x avec U_1 pour essayer de savoir dans quelle mesure chacune d'elles participe à sa constitution.

On peut projeter les variables x et y dans des plans engendrés par les variables canoniques de W_1 et dans des plans engendrés par celles de W_2, ce qui permet de visualiser les résultats d'une manière un peu semblable à ce qui a été fait pour l'analyse factorielle.

L'analyse canonique est ainsi une généralisation de la régression et concerne des ensembles de variables quantitatives qu'on a quelque raison de partitionner en deux sous-ensembles. On montre que l'analyse des correspondances et l'analyse discriminante peuvent être considérées comme des cas particuliers de l'analyse canonique.

Classifications automatiques

Ce sont les naturalistes qui, à la suite de Linné (1707-1778), ont établi avec l'inventaire systématique du règne animal et du règne végétal la plus célèbre des classifications. Leurs groupements en règnes, embranchements, ordres, familles, tribus, genres, espèces fournissent un exemple particulièrement riche de classification hiérarchique. L'avènement des ordinateurs, l'élaboration de critères de classification et la mise au point d'algorithmes ont permis d'étendre ces méthodes à des domaines variés, en sciences exactes et en sciences humaines.

Toute classification a pour but de déterminer des regroupements sur un ensemble de données afin d'en obtenir une perception simplifiée mais globale. En termes mathématiques, cela revient à construire une partition de l'ensemble considéré. On distinguera la notion de **classification** de celle de **classement** qui consiste à placer un élément quelconque dans une classe déjà déterminée. Les problèmes abordés peuvent être de nature extrêmement diverse : ainsi on peut chercher à regrouper des régions suivant leurs activités économiques, élaborer une nomenclature de ces activités, analyser des réponses à un questionnaire, étudier la répartition de la faune ou de la flore d'une certaine région, etc.

Il existe plusieurs méthodes de résolution du problème de classification. On peut procéder :

— *par regroupements successifs des constituants :* on suppose au stade initial que tous les individus appartiennent à des classes différentes; à chaque pas, on regroupe certains d'entre eux dans une même classe; les partitions deviennent de moins en moins fines *(classification ascendante) ;*

— *par fractionnements successifs des classes :* au stade initial, tous les individus sont dans une même classe que l'on sépare en deux ou plusieurs sous-ensembles disjoints, puis on réitère le procédé *(classification descendante).*

Classification hiérarchique ascendante

La **première étape** consiste à définir une mesure de proximité sur l'ensemble des individus à classer. On appelle *mesure de proximité* la donnée d'un tableau symétrique $n \times n$ où la case (i, j) indique la ressemblance des individus i et j. Le choix de cette mesure est un des points les plus délicats et nécessite une bonne connaissance du domaine étudié. Parmi les mesures de proximité, on peut citer :

- *l'indice de similarité s*

s est une application $E \times E \to \mathbb{R}^+$ telle que

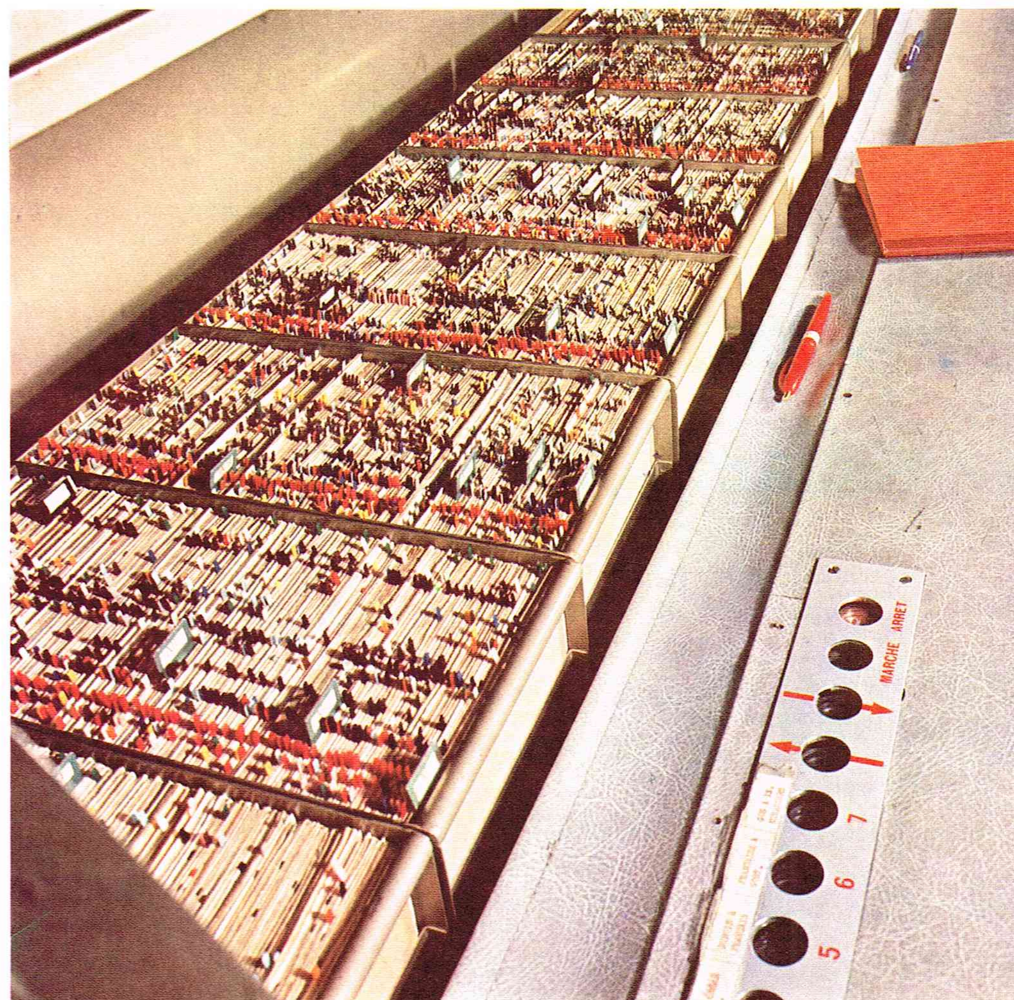
$s(x, y) = s(y, x) \quad \forall x, y \in E$
$s(x, x) = s(y, y) \geqslant s(x, y) \quad \forall x, y \in E$

Lorsque pour chaque individu on relève la présence ou l'absence de p caractères (variables), on peut poser :

s = nombre de caractères possédés à la fois par i et j
t = nombre de caractères non possédés à la fois par i et j
u = nombre de caractères possédés par i et non par j
v = nombre de caractères possédés par j et non par i

Les indices suivants ont été utilisés par les taxinomistes :

indice de Sokal et Nichener (1958) : $\dfrac{s+t}{p}$



Ciccione - Rapho

▲ ***Casiers à fiches.***

indice de Rogers et Tanimoto (1960) : $\dfrac{s+t}{s+t+u}$

indice de Sokal et Sneath : $\dfrac{s+t}{u+v}$

- *l'indice de dissimilarité*

$d : E \times E \to \mathbb{R}^+$ vérifiant
(a) $d(x, y) = d(y, x) \quad \forall x, y \in E$
(b) $d(x, x) = 0$

- *l'indice de distance*

d vérifie **(a)**, **(b)** et **(c)**
(c) $d(x, y) = 0 \Rightarrow x = y \quad \forall x, y \in E$

- *la distance*

d vérifie **(a)**, **(b)**, **(c)** et **(d)**
(d) $d(x, y) \leqslant d(x, y) + d(y, z) \quad \forall x, y, z \in E$

- *la distance ultramétrique*

d vérifie **(a)**, **(b)**, **(c)** et **(e)**
(e) $d(x, y) \leqslant \text{Max}\,(d(x, z), d(z, y)) \quad \forall x, y, z \in E$

On démontre que **(d)** $\Rightarrow$ **(e)** et que si d est une distance ultramétrique, alors tout triangle est isocèle, et lorsque deux boules ont un point commun, l'une est incluse dans l'autre. L'indice de similarité est transformé en indice de dissimilarité par $d(x, y) = s(x, x) - s(x, y)$.

A toute partition d'un ensemble on peut associer une distance ultramétrique en posant :

$\forall i \in E\ d(i, i) = 0$, $d(i, j) = 1$ pour tous les couples i, j appartenant à deux classes distinctes.
$d(i, j) = l\ (0 < l < 1)$ si i et j appartiennent à la même classe.

Ainsi la mesure de proximité peut être une distance et même une distance ultramétrique, mais ce choix n'est pas toujours possible; cela dépend essentiellement de la nature des données.

Dans la **deuxième étape,** on est amené à agglomérer des classes ; il faudra donc définir une proximité entre parties de E (parties éventuellement réduites à un élément). Si A et B désignent deux parties non vides de E et d, l'indice de proximité sur E, on peut poser :

— δ_{max} (A, B) = Max $\{d(x, y) \mid x \in A, y \in B\}$
— δ_{min} (A, B) = Min $\{d(x, y) \mid x \in A, y \in B\}$
— $\delta_{max\ min}$ (A, B) =
$$\text{Max}\left\{\underset{x \in A}{\text{Max}}\ \underset{y \in B}{\text{Min}}\ d(x, y),\ \underset{y \in B}{\text{Max}}\ \underset{x \in A}{\text{Min}}\ d(x, y)\right\}$$
— si E est un espace métrique
$$\delta_g\ (A, B) = d\ (G_A, G_B)$$

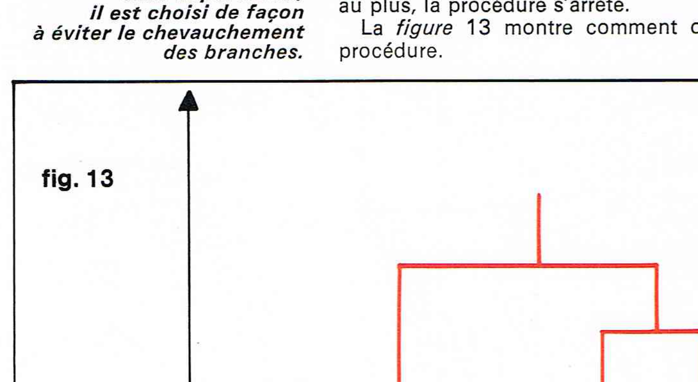
G_A, G_B désignant les centres de gravité de A et B.

Les δ sont des indices de proximité mais en général pas des distances.

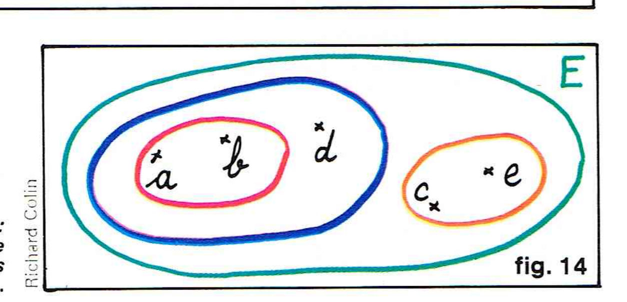
Là encore il faut choisir l'indice qui s'adapte le mieux au phénomène étudié. Lorsque le choix ne s'impose pas, il est prudent de se préoccuper de la stabilité des résultats obtenus.

Les **étapes suivantes** consistent à regrouper pas à pas les classes les plus proches. Au bout de $(n-1)$ pas au plus, la procédure s'arrête.

La *figure* 13 montre comment on peut visualiser la procédure.



▼ ***Figure 13 :*** *l'ordre des individus* **a, b, ... e** *sur l'axe est sans importance ; il est choisi de façon à éviter le chevauchement des branches.*



▶ ***Figure 14 :*** *visualisation, sous forme de hiérarchie de parties de E, de la procédure.*

Ici quatre pas ont été nécessaires. L'arbre ainsi construit permet de découvrir les partitions successives apparues à chaque pas. La partition initiale en classes à un élément $\{a\}$, $\{b\}$, $\{c\}$, $\{d\}$, $\{e\}$ devient à l'issue du premier pas $\{a, b\}$, $\{c\}$, $\{d\}$, $\{e\}$ puis $\{a, b, d\}$, $\{c\}$, $\{e\}$ et enfin $\{a, b, d\}$, $\{c, e\}$ avant le regroupement final. Les partitions réalisées sont de moins en moins fines (on dit qu'une partition P est plus fine qu'une partition P', si tout élément de P est élément de P').

Hiérarchie stratifiée et ordonnance

La procédure aurait pu aussi bien être représentée (*fig.* 14) sous forme de hiérarchie de parties de E .

H est une hiérarchie de parties de E si :
— H est un sous-ensemble de parties de E ($H \subset \mathfrak{P}(E)$)
— $\{x\} \in H \quad \forall x \in E$
— $E \in H$
— $A \cap B \in \{A, B, \varnothing\} \quad \forall A \in H, \quad \forall B \in H$

ce qui traduit que deux éléments de H sont disjoints ou sinon l'un est inclus dans l'autre.

La donnée d'une hiérarchie H ne permet pas de savoir dans quel ordre se sont effectués les regroupements. Par exemple, c et e peuvent avoir été réunis avant a, b et d. Pour préciser que l'ordre des groupements est connu, on dit que l'on a une *hiérarchie stratifiée*, alors dans l'arbre de classification on connaît l'ordre de formation des nœuds, ce que l'on visualise en les traçant d'autant plus haut qu'ils correspondent à des regroupements plus tardifs. Les partitions compatibles avec une hiérarchie stratifiée sont celles que l'on obtient en coupant par des horizontales l'arbre de classification associé.

La *figure* 15 montre cinq classifications compatibles avec l'arbre de la *figure* 13.

L'examen de cet arbre permet de dire en outre si telle paire d'éléments de E a été réunie ou non avant telle autre. Une hiérarchie stratifiée sur E induit un ordre sur les paires d'éléments de E, un tel ordre est appelé *ordonnance* sur E.

Pour l'exemple choisi on aura :

$\{a, a\} = \{b, b\} = \{c, c\} = \{d, d\} = \{e, e\} < \{a, b\} < \{a, d\} = \{b, d\} < \{c, e\} < \{a, c\} = \{a, e\} = \{b, c\} = \{b, e\} = \{d, c\} = \{d, e\}$.

Hiérarchie indicée et ultramétrique

On améliore encore les résultats précédents en adoptant une mesure sur l'axe vertical de l'arbre : à chaque nœud sera associé un indice d'autant plus grand que le nœud sera plus élevé sur l'axe vertical, on dira alors que l'on a une *hiérarchie indicée*.

Une hiérarchie indicée est une hiérarchie stratifiée sur laquelle on a défini une fonction réelle V qui est telle que :
$\forall x \in I \quad V(x) = 0$ (les classes à un seul élément ont un indice nul).
$A \subset B \Rightarrow V(A) \leqslant V(B)$ [l'indice voit sa valeur augmenter à mesure que l'on s'élève dans l'arbre].

On démontre que toute hiérarchie indicée induit sur E une distance ultramétrique en prenant :

$$d\ (i, j) = \text{Min}\,\{V(A) \mid A \in H, i \in A, j \in A\} \qquad \forall i, j \in E$$

$d\ (i, j)$ est l'indice de la plus petite classe contenant à la

fig. 15

=> {a, b, c, d, e}
=> {{a, b, d}, {c, e}}
=> {{a, b, d}, {c}, {e}}
=> {{a, b}, {d}, {c}, {e}}
=> {{a}, {b}, {c}, {d}, {e}}

a b d c e

Richard Colin

▶ ***Figure 15 :*** *cinq classifications compatibles avec l'arbre de la figure 13.*

fois i et j ; c'est encore l'indice du premier nœud rassemblant à la fois i et j. Réciproquement on démontre qu'à toute distance ultramétrique correspond une hiérarchie indicée équivalente (induisant la même ordonnance) à la famille des hiérarchies indicées associées à une même hiérarchie stratifiée.

Ceci montre tout l'intérêt de la distance ultramétrique. Pour construire un arbre de classification, on cherche à approcher l'indice de proximité par une distance ultramétrique. On démontre que parmi toutes les ultramétriques inférieures à d (indice de proximité) il en existe une supérieure à toutes les autres (ultramétrique sous dominante de d) et que c'est la plus proche de d pour n'importe quelle distance entre indices de dissimilarité. Divers algorithmes sont utilisés.

Procédés opérant par dichotomies

L'objectif est toujours d'obtenir une partition de E en classes qui regroupent les éléments qui se ressemblent le plus, mais cette fois en opérant par fractionnements successifs de l'ensemble E.

Partitionner E en deux classes équivaut à définir une application booléenne sur E (application qui prend la valeur 1 pour les éléments d'une classe et la valeur 0 pour les autres). Il paraîtra logique de définir une « distance » (au sens indice de proximité) entre classes puis de choisir parmi les $2^{n-1} - 1$ applications booléennes celle qui maximise la distance entre deux classes. Le même procédé permet de fractionner les deux classes obtenues, etc. (*fig.* 16 et 17). On obtient ainsi une *segmentation* de l'ensemble E.

Un des inconvénients consiste en ce que deux individus séparés lors de la première dichotomie seront très éloignés alors qu'ils diffèrent peut-être seulement par le premier caractère ayant servi à établir la séparation.

Là encore divers algorithmes ont été proposés, trop nombreux pour que nous puissions en rendre compte ici.

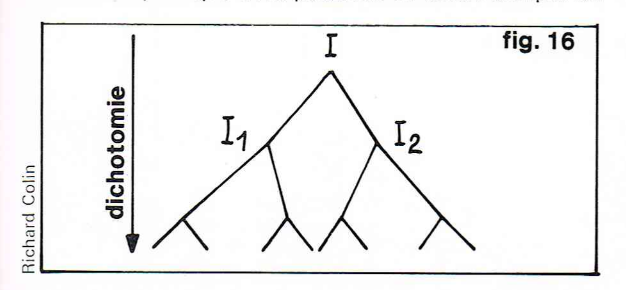


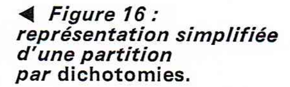
Tableau récapitulatif des différentes méthodes utilisées en analyse des données.

But	Méthodes	Nature des variables	
But descriptif	Analyse en composantes principales	Variables quantitatives	
	Analyse des correspondances	Variables quantitatives ou qualitatives	
		Variables à expliquer	Variables explicatives
But explicatif ou prévisionnel	Analyse discriminante	p quantitatives	1 qualitative
	Régression	1 quantitative	p quantitatives
	Analyse canonique	p quantitatives	q quantitatives
But descriptif	Classification	Variables qualitatives et/ou quantitatives	

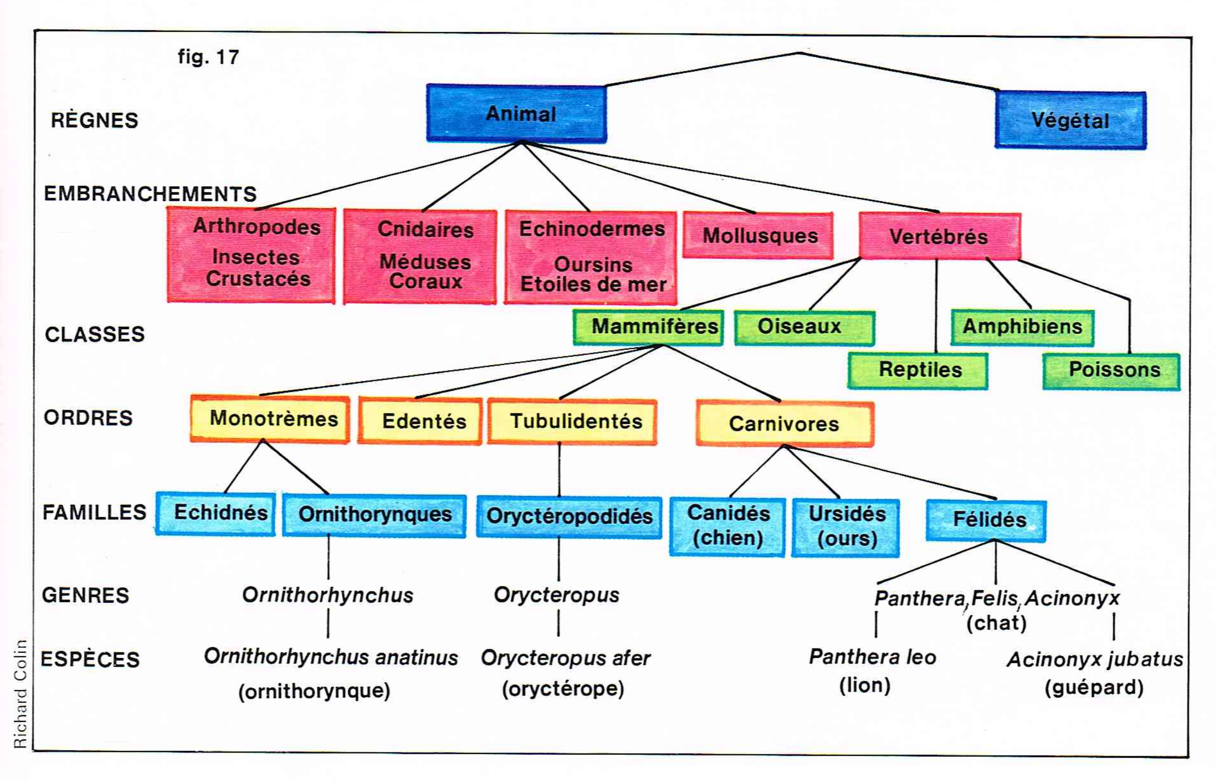
▲ *Tableau récapitulatif des différentes méthodes utilisées en analyse des données.*

BIBLIOGRAPHIE

ANDERSON T. W., *An Introduction to Multivariate Statistical Analysis*, Wiley and Son, New York, 1958. - BENZECRI J.-P., *l'Analyse des données : tome I, la Taxinomie, tome II, l'Analyse des correspondances*, Dunod, Paris, 1973. - KANE E. J., *Statistique économique et Économétrie*, Colin, Paris, 1971. - LEBART L. et FENELON J. P., *Statistique et informatique appliquées*, Dunod, Paris, 1971. - LINNIK Y. V., *la Méthode des moindres carrés*, Dunod, Paris. - MALINVAUD E., *Méthodes statistiques de l'économétrie*, Dunod, Paris, 1969. - MORONEY M. J., *Comprendre la statistique*, Marabout Université, Paris, 1970. - ROMEDER J. M., *Analyse discriminante*, Masson, Paris, 1971. - VANGREVELINGHE G., *Économétrie*, Hermann, Paris, 1973. - Publication du Centre d'études économiques d'entreprise, sous la direction de G. MORLAT : *Analyse de données multidimensionnelles* (3 tomes), Paris, 1971.



◄ *Figure 16 : représentation simplifiée d'une partition par* **dichotomies.**



◄ *Figure 17 : illustration tirée du tome I* la Taxinomie, *2e édition, 1976, de l'ouvrage* l'Analyse des données *par J.-P. Benzecri et coll. (Dunod).*



Photo Research International

CALCUL NUMÉRIQUE

De tous temps l'homme a cherché à calculer, mesurer, évaluer les phénomènes auxquels il s'est trouvé confronté. Les géomètres de l'antique Égypte étaient chargés de mesurer la surface des champs que périodiquement le Nil recouvrait d'un limon salvateur et bienfaisant. Pour cela il leur fallait non seulement avoir imaginé un *instrument de mesure* (chaîne d'arpenteur, etc.), mais encore disposer d'une *méthode* permettant de calculer la surface d'un champ quelle qu'en soit la forme. Cette méthode ne pouvait exister que fondée sur un support théorique solide c'est-à-dire sur la géométrie. La géométrie leur démontrait que tout polygone pouvait se décomposer en triangles, la méthode consistait à déterminer la formule donnant la surface d'un triangle; restait la mesure proprement dite qui, se réduisant ainsi à une succession d'opérations élémentaires de simple manutention, pouvait dès lors être confiée à quelques travailleurs d'outre-Nil.

Le calcul numérique était né de cette confrontation entre un corps de théorie (le cas échéant créé pour la circonstance), et un problème réel. L'avènement de l'informatique et l'irruption des ordinateurs dans le monde du XXe siècle ont permis au calcul numérique de connaître un développement sans précédent. Ce développement est à l'évidence commandé par le nombre croissant de problèmes qui, aujourd'hui, dans de multiples domaines (physique, technologie, économie, recherche opérationnelle, etc.), suggèrent une formalisation mathématique et, subséquemment, une résolution numérique aussi rapide et précise que possible.

Songeons par exemple au programme « Apollo » et notamment à la première expédition sur la Lune de juillet 1969 : la trajectoire du vaisseau spatial, sa vitesse à chaque instant, l'angle de sa rentrée dans l'atmosphère, etc., sont autant de grandeurs mesurables qui sont obtenues au terme de calculs mathématiques très élaborés, et qu'il doit être possible de faire de manière instantanée si les circonstances (même les plus imprévues) l'exigent. Tous les paramètres sont calculés à partir d'un *modèle* mathématique global décrivant toutes les relations qui existent entre eux, ainsi que les équations qui les régissent. Grâce à ce modèle, il est possible de prévoir l'évolution du système.

L'homme de la rue a quelque peine à imaginer la complexité d'un tel modèle, mais il peut percevoir que la démarche des techniciens et ingénieurs de la N. A. S. A. reste au fond très semblable à celle des géomètres égyptiens : le passage d'un problème réel à sa résolution pratique se fait toujours en trois étapes : la *mise au point du modèle* mathématique décrivant (au mieux) le problème réel (ici le modèle doit faire appel aussi bien à l'astronomie, à la balistique, à l'électronique, qu'aux mathématiques), la *mise au point de la méthode* (ou *algorithme*) pour conduire les calculs, enfin l'*élaboration du programme informatique*, c'est-à-dire la traduction de l'algorithme en un langage accessible au calculateur électronique. La chaîne d'arpenteur est devenue ordinateur.

Il est certain que ces trois étapes restent étroitement complémentaires : les possibilités de l'informatique astreignent tout algorithme à certaines contraintes que l'on ne peut négliger (coût du temps de passage, capacité de mémoire, etc.). De même les progrès de la recherche mathématique sont indispensables pour la résolution de nouveaux problèmes (notamment dans le domaine de l'analyse numérique, c'est-à-dire l'étude des équations aux dérivées partielles). C'est précisément cette deuxième étape qui est le champ d'action du calcul numérique. Le calcul numérique est donc la science de l'algorithme.

Définition d'un algorithme

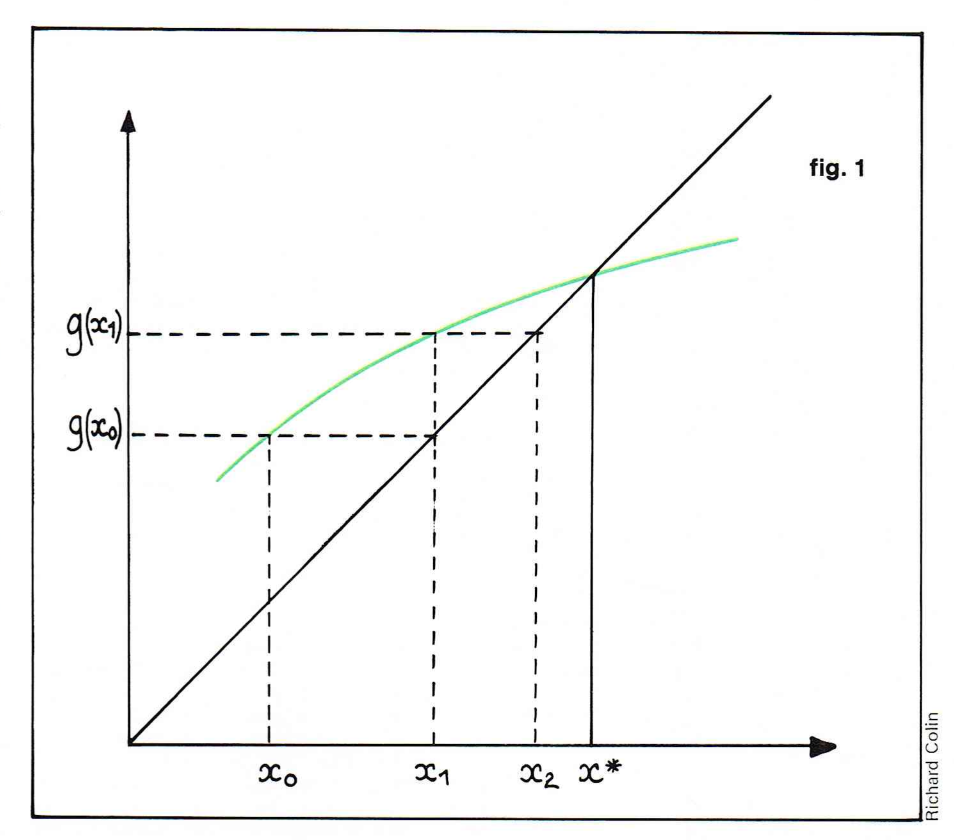
Un algorithme est une suite d'opérations élémentaires (en général algébriques) qui conduit à la résolution (exacte ou approchée) d'un problème. Un algorithme est donc formé d'une succession d'*étapes* (ou *itérations*) qui s'enchaînent de manière logique.

La solution sera exacte ou approchée suivant la nature du problème posé. Un calcul est dit exact lorsque la solution recherchée est donnée de manière analytique, c'est-à-dire à l'aide d'une formule composée d'opérations algébriques connues (par exemple : la recherche des racines d'une équation du deuxième degré, l'inversion d'une matrice, etc.). Le calcul approché répond, quant à lui, à l'émergence de problèmes plus complexes dont il n'existe pas de solution analytique. On cherche donc à se rapprocher de la solution théorique avec une précision choisie à l'aide d'un *algorithme convergent* mais qui ne l'atteindrait qu'en un nombre infini d'itérations (recherche de racine d'une équation de degré > 2, calcul d'intégrale par interpolation, résolution d'équations différentielles non linéaires ou aux dérivées partielles, recherche d'optimum, etc.).

Exemple : donnons un exemple simple qui, illustrant ce type de méthodes, donne une idée des difficultés qui peuvent surgir. Considérons l'équation à une inconnue :

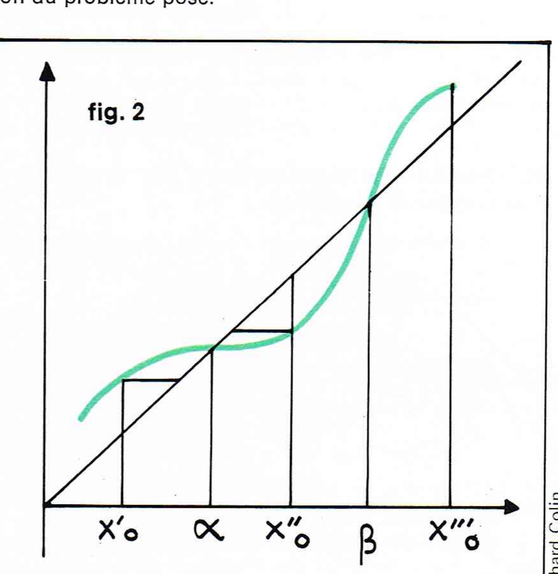
(1) $$g(x) = x$$

où g est une fonction numérique. Géométriquement, la solution est l'ensemble des points d'intersection de la courbe $y = g(x)$ avec la première bissectrice. Partant d'un point x_0, il est facile de construire le point $x_1 = g(x_0)$ puis $x_2 = g(x_1)$, etc. Sur la *figure 1* il est clair que l'on définit ainsi une suite x_n à l'aide de la formule récurrente $x_{n+1} = g(x_n)$ et que cette suite converge vers x^*, solution du problème posé.



▲ *Figure 1 : un* **algorithme** *est une suite d'opérations élémentaires qui conduit à la résolution d'un problème; la* **formule récurrente** $x_{n+1} = g(x_n)$ *définit une suite* x_n *qui converge vers* x^*.

◄ *Page ci-contre, une des salles du centre de contrôle de la N. A. S. A.*



◄ *Figure 2 : en prenant comme valeur initiale* x'_0 *ou* x''_0*, l'algorithme converge vers* α *mais il n'est pas possible de trouver une valeur initiale qui le fasse converger vers* β *qui est également solution de* **g (x) = x**.

L'algorithme ainsi décrit est un *processus itératif* (ou à approximations successives). A chaque étape de l'algorithme, on se rapproche de la solution, on détermine le résultat de l'étape $n + 1$ à l'aide de celui de l'étape n.

Il est aisé de voir que si l'algorithme converge et si, en outre, la fonction est continue, alors la limite x^* de la suite x_n est solution de l'équation. Mais la convergence elle-même suppose que la fonction y vérifie certaines hypothèses. De surcroît, cette convergence peut dépendre de la valeur initiale choisie.

Ainsi, dans le cas représenté sur la *figure* 2, en prenant comme valeur initiale de x'_0 ou x''_0, l'algorithme converge vers α mais il n'est pas possible de trouver une valeur initiale qui le fasse converger vers β qui est également solution de l'équation **(1)**; de plus, en x'''_0 le processus diverge ! La convergence d'un algorithme va donc dépendre des propriétés mathématiques des concepts qui interviennent dans sa définition. Mais d'un point de vue pratique, il est important de pouvoir connaître en outre la *vitesse de convergence* dudit algorithme. La vitesse de convergence peut par exemple être mesurée par le nombre d'itérations nécessaire pour passer d'une valeur initiale x_0 à une valeur contenue dans un voisinage de rayon ε de la solution x^*. Si la vitesse de convergence est trop faible, il pourra être nécessaire de changer l'algorithme ou de modifier celui-ci afin de diminuer le temps de passage sur l'ordinateur.

Nous venons de voir un exemple simple où la solution était un nombre. Dans certains problèmes plus généraux, x représente un vecteur de $\mathbb{R}^p$ ou même une fonction. Ainsi la résolution d'une équation différentielle peut être approchée par un processus (*méthode* dite *d'Euler* par exemple). Pour parler de convergence d'un tel algorithme, il faut qu'au préalable on ait défini non seulement l'*espace fonctionnel* dans lequel on cherche à s'approcher de la solution mais encore que l'on ait muni cet espace d'une norme ou d'une distance (c'est donc dans le cadre de l'analyse fonctionnelle [voir *Analyse fonctionnelle*] qu'on pourra replacer ce type de problèmes).

Erreur

Considérons un processus itératif du type de celui donné en exemple, où x désigne plus généralement un vecteur ou une fonction, c'est-à-dire mathématiquement un élément d'un espace vectoriel topologique (par exemple normé), la suite x_n converge donc vers la solution x^*, autrement dit : $\lim_{n \to \infty} \| x_n - x^* \| = 0$

Supposons que l'on arrête le processus à l'itération N, alors l'*erreur absolue* commise en prenant x_N comme solution approchée est :

$$e_n = \| x_N - x^* \| ;$$

l'*erreur relative* est alors donnée par

$$e'_n = \frac{\| x_n - x^* \|}{\| x^* \|}$$

Ne connaissant pas x^*, il s'agit de trouver des majorations de e_n, e'_n (ne dépendant pas de x^*) de manière à évaluer une borne supérieure de l'erreur (absolue ou relative) commise en arrêtant l'algorithme à l'étape N. Là encore, les majorations possibles dépendent des propriétés mathématiques de l'algorithme considéré.

Méthodes de résolution d'un système d'équations linéaires

Considérons le système de n équations à n inconnues suivant :

$$\begin{cases} a_{11} x_1 + \ldots + a_{1n} x_n = b_1 \\ a_{21} x_1 + \ldots + a_{2n} x_n = b_2 \\ a_{n1} x_1 + \ldots + a_{nn} x_n = b_n \end{cases}$$

En définissant la matrice A et les vecteurs $\vec{x}$ et $\vec{b}$ par :

$$A \begin{bmatrix} a_{11} \ldots & a_{1n} \\ & \vdots \\ a_{n1} & a_{nn} \end{bmatrix}, \quad \vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \quad \vec{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}.$$

On peut mettre le système sous forme matricielle :

$$A\vec{x} = \vec{b}.$$

D'après ce qui a été vu en algèbre linéaire, cette équation matricielle a une solution unique $\vec{x}^*$ si, et seulement si, le déterminant de la matrice A est non nul (*système* dit *de Cramer*); la solution cherchée est alors obtenue en inversant la matrice A : $\vec{x}^* = A^{-1} b$.

Une méthode d'inversion de matrice a déjà été donnée (voir *Algèbre linéaire*) : elle oblige à calculer un grand nombre de déterminants (n^2 déterminants d'ordre $n - 1$) ce qui rend le calcul particulièrement fastidieux lorsque n est grand. D'un point de vue pratique, il est préférable d'utiliser d'autres méthodes plus opérationnelles. Il en est de deux types :

— les **méthodes exactes** qui sont des algorithmes finis de calcul de solutions (méthode de Gauss, méthode du pivot) ;

— les **méthodes itératives**, analogues à l'exemple développé dans le chapitre précédent, qui permettent d'obtenir les solutions, avec la précision voulue, à l'aide de processus convergents infinis (méthode des approximations successives, méthode de Seidel, méthode de relaxation). Les approximations (erreurs d'arrondis, etc.) inévitables font que même les résultats des méthodes exactes sont approchés; l'utilisation des méthodes itératives donne lieu, de surcroît, à l'erreur inhérente à la méthode. Notons d'ores et déjà que l'efficacité des méthodes itératives va dépendre du bon choix de la valeur initiale et de la rapidité de la convergence du processus.

Soit donc à résoudre le système :

$$\begin{cases} 2x_1 + 3x_2 + x_3 = 4 & \textbf{(a)} \\ x_1 - x_2 - 2x_3 = 1 & \textbf{(b)} \\ x_1 + 4x_2 + x_3 = 1 & \textbf{(c)} \end{cases}$$

Étape 1 : on divise l'équation **(a)** par le coefficient de x_1 dans cette équation. D'où l'équation **(a')**; on élimine x_1 des équations **(b)** et **(c)** en retranchant à chacune d'elles l'équation **(a')** :

$$x_1 + \frac{3}{2} x_2 + \frac{1}{2} x_3 = 2 \quad \textbf{(a')}$$

$$-\frac{1}{2} x_2 - \frac{5}{2} x_3 = -1 \quad \textbf{(b')}$$

$$\frac{5}{2} x_2 + \frac{1}{2} x_3 = -1 \quad \textbf{(c')}$$

Étape 2 : on conserve l'équation **(a')**; on élimine x_2 de l'équation **(c')** en divisant l'équation **(b')** par le coefficient de x_2 dans cette équation; d'où l'équation **(b")**. On élimine x_2 de **(c')** en retranchant de cette équation l'équation **(b")** multipliée par $\frac{5}{2}$:

$$x_1 + \frac{3}{2} x_2 + \frac{1}{2} x_3 = 2 \quad \textbf{(a')}$$

$$x_2 + 5x_3 = 2 \quad \textbf{(b")}$$

$$-12x_3 = -6 \quad \textbf{(c")}$$

Étape 3 : ce système est sous forme diagonale; il admet pour solution :

$$x_1 = \frac{1}{2}$$

$$x_2 = -\frac{1}{2}$$

$$x_3 = \frac{5}{2}$$

▶ ***Tableau synoptique du principe de la méthode des éliminations successives.***

Méthode des éliminations successives ou méthode de Gauss

La méthode consiste à remarquer que la résolution d'un système linéaire mis sous forme diagonale est facile à résoudre.

Exemple :

$$\begin{cases} 2\,x_1 + x_2 - x_3 = 1 & (2) \\ 3\,x_2 + x_3 = 4 & (3) \\ 3\,x_3 = 3 & (4) \end{cases}$$

De l'équation **(4)** on tire $x_3 = 1$ que l'on reporte dans l'équation **(3)** pour obtenir la valeur de $x_2 = 1$; l'équation **(2)** donne alors celle de $x_1 \left(= \frac{1}{2}\right)$. L'*algorithme de Gauss* consiste à mettre le système à résoudre sous forme diagonale en procédant à certaines combinaisons linéaires des équations, de manière à éliminer une ou plusieurs variables. Ces combinaisons successives s'effectuent sur les coefficients du système ; à chaque étape, dans la pratique, on se contente d'écrire les coefficients, sous forme de *tableaux*.

Voyons le principe de la méthode sur un exemple que nous résoudrons de manière synoptique à gauche par l'élimination successive des variables, à droite par la manipulation des tableaux des coefficients (ci-dessous).

Tableau 0 :

$\boxed{2}$	3	1	4
1	1	-2	1
1	4	1	1

Tableau 1 : on multiplie la première ligne par $\frac{1}{2}$ que l'on soustrait alors des deux autres.

2 s'appelle *pivot ;* on l'encadre dans le tableau 0 :

1	$\frac{3}{2}$	$\frac{1}{2}$	2
0	$\boxed{-\frac{1}{2}}$	$-\frac{5}{2}$	-1
0	$\frac{5}{2}$	$\frac{1}{2}$	-1

Tableau 2 : on ne modifie pas la première ligne. On multiplie la deuxième par -2. Cette ligne ainsi transformée est multipliée par $-\frac{5}{2}$ pour être ajoutée à la troisième de manière à faire apparaître un 0 à la place du $\frac{5}{2}$ du tableau 1 .

$\boxed{-\frac{1}{2}}$ est le deuxième pivot :

1	$\frac{3}{2}$	$\frac{1}{2}$	2
0	1	5	2
0	0	$\boxed{-12}$	-6

$\boxed{-12}$ est le troisième pivot ou en multipliant la dernière ligne par $-\frac{1}{12}$.

Tableau 3 :

1	$\frac{3}{2}$	$\frac{1}{2}$	2
0	1	5	2
0	0	1	$\frac{1}{2}$

On voit donc déjà apparaître la notion de *pivot*. La méthode du pivot que nous présentons maintenant est donc une forme améliorée de la méthode de Gauss.

Méthode du pivot

La méthode du pivot consiste à mettre le système sous forme *diagonale* en effectuant des combinaisons linéaires des lignes analogues à celles qui interviennent dans la méthode de Gauss.

Exemple : nous reprenons l'exemple précédent en utilisant la disposition pratique sous forme de tableaux que nous avons introduits :

Tableau 0 :

$\boxed{2}$	3	1	4
1	1	-2	1
1	4	1	1

Le premier pivot étant choisi (ici le coefficient de x_1 dans la première équation), la méthode consiste à procéder à des combinaisons de lignes de manière à faire apparaître un $\boxed{1}$ à la place du pivot et des 0 à la place des autres coefficients de la première ligne. Le *tableau 1* s'obtient donc de la même façon que dans l'exemple précédent :

Tableau 1 :

1	$\frac{3}{2}$	$\frac{1}{2}$	1
0	$\boxed{-\frac{1}{2}}$	$-\frac{5}{2}$	-1
0	$\frac{5}{2}$	$\frac{1}{2}$	-1

On prend comme deuxième pivot le coefficient de x_2 dans la deuxième équation (on aurait tout aussi bien pu prendre la troisième équation). On multiplie la deuxième ligne par -2 puis par $-\frac{3}{2}$ $\left(\text{respectivement } -\frac{5}{2}\right)$ pour l'ajouter ainsi transformée à la première (respectivement troisième) ligne de manière à faire apparaître des 0 en lieu et place de $\frac{3}{2}$ et $\frac{5}{2}$; on obtient ainsi le *tableau 2 :*

Tableau 2 :

1	0	-7	-1
0	1	5	2
0	0	$\boxed{-12}$	-6

En prenant comme troisième pivot $\boxed{-12}$ et en procédant de la même façon, on aboutit au *tableau 3 :*

Tableau 3 :

1	0	0	$\frac{5}{2}$
0	1	0	$-\frac{1}{2}$
0	0	1	$\frac{1}{2}$

La solution du système est là encore :

$$x_1 = \frac{5}{2}, \quad x_2 = -\frac{1}{2}, \quad x_3 = \frac{1}{2}.$$

Exemple : cette méthode est également utilisée pour calculer l'*inverse d'une matrice donnée*. Donnons-en le principe sans en fournir la justification théorique. Sur l'exemple précédent, il suffit d'appliquer la méthode du pivot au tableau suivant :

2	3	1	1	0	0
1	1	-2	0	1	0
1	4	1	0	0	1

En prenant successivement les éléments diagonaux du bloc de gauche comme pivots, on obtient au bout de la troisième étape :

$$\begin{array}{ccc|ccc} 1 & 0 & 0 & \frac{3}{4} & \frac{1}{12} & -\frac{7}{12} \\ 0 & 1 & 0 & -\frac{1}{4} & \frac{1}{12} & \frac{5}{12} \\ 0 & 0 & 1 & \frac{1}{4} & -\frac{5}{12} & -\frac{1}{12} \end{array}$$

Le bloc de droite du tableau ainsi obtenu est la matrice inverse de la matrice cherchée.

Avantages comparés des méthodes proposées

Du point de vue de l'utilisation, la meilleure méthode est celle qui donne lieu au plus petit nombre d'opérations élémentaires (addition, multiplication). Il est possible de calculer le nombre N d'opérations requises dans l'application des 3 méthodes, en fonction du nombre n d'inconnues.

Méthode de Cramer (déterminants)	$N_c = (n^2 + n)\, n\,! - 1$
Méthode de Gauss (élimination successive)	$N_g = 2\, n\, (n + 1)\, (n + 2) + [n\, (n - 1)]$
Méthode du pivot	$N_p = 4\, n^3 - 2\, n$

▲ *Avantages comparés des 3 méthodes proposées en fonction du nombre n d'inconnues.*

Pour $n \geqslant 4$, on a les inégalités $N_g \simeq N_p < N_c$. Ainsi la méthode de Gauss semble préférable pour la résolution de gros systèmes linéaires : supposons que l'on ait à résoudre un système de 10 équations et que l'on utilise un calculateur qui effectue 10^4 opérations par seconde. Le temps T nécessaire pour résoudre le système est dans les 3 cas : $T_c \simeq$ 24 heures ! T_g 2 1/10 sec. $T_p \simeq$ 4 1/10 sec. On comprend pourquoi la méthode de Cramer n'a aujourd'hui qu'un intérêt académique.

Méthode des approximations successives

Lorsque le nombre des inconnues devient trop important, l'application des méthodes exactes que nous venons de voir s'avère bien trop compliquée. On leur préfère alors des méthodes numériques approchées. L'une d'elles est la méthode des approximations successives :

Considérons le système linéaire **(5)** : $A\vec{x} = \vec{b}$. Il est toujours possible de trouver une matrice H et un vecteur $\vec{k}$ tels que **(5)** devienne **(6)** : $\vec{x} = H\vec{x} + \vec{k}$. Si A est la matrice (a_{ij}), on vérifiera sans peine que la matrice H est égale à :

$$H = \begin{bmatrix} 0 - \frac{a_{12}}{a_{11}} \cdots - \frac{a_{1n}}{a_{11}} \\ \\ 0 \\ \\ - \frac{a_{n1}}{a_{nn}} \qquad 0 \end{bmatrix} \quad \text{et} \quad \vec{k} = \begin{bmatrix} \frac{b_1}{a_{11}} \\ \\ \\ \\ \frac{b_n}{a_{nn}} \end{bmatrix}$$

Cette transformation est toujours possible si la matrice A est inversible. En effet, on peut supposer qu'alors les éléments de la diagonale sont tous $\neq 0$, quitte à modifier l'ordre des lignes ou des colonnes. On choisit une valeur initiale

$\vec{x}^0 = [x_1^0, \ldots, x_n^0]$ (en général on prend $\vec{x}^0 = (0, \ldots 0, \ldots 0)$). On calculera x^1 par la formule $\vec{x}^1 = \vec{H}x^0 + \vec{k}$. A la i-ième itération, on aura $\vec{x}^{i+1} = \vec{H}x^i + \vec{k}$. Si le processus converge vers une limite $\vec{x}^*$, alors il est clair que $\vec{x}^*$ est solution du système considéré. Pour que l'algorithme converge, il faut et il suffit que toutes les valeurs propres de la matrice H soient en modules inférieures à 1. Cette condition est assurée lorsque les coefficients diagonaux de la matrice A sont, en module, plus grands que la somme des autres éléments de leur ligne (et leur colonne) :

$$|\alpha_{ii}| > \sum_{j \neq i} \alpha_{ij}$$

Exemple : soit à résoudre le système :

$$\begin{aligned} 10\, x_1 - 2\, x_2 - 2\, x_3 &= 6 \\ -x_1 + 10\, x_2 - 2\, x_3 &= 7 \\ -x_1 - x_2 + 10\, x_3 &= 8 \end{aligned}$$

le système s'écrit sous la forme **(6)** :

(6) k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$
0	0,6	0,7	0,8
1	0,88	0,92	0,93
2	0,97	0,974	0,98
3	0,991 8	0,993	0,994 4

Le processus converge vers la solution (évidente) $x_1^* = x_2^* = x_3^* = 1$, sans l'atteindre en un nombre fini d'itérations. Cependant, il faut noter que ce processus itératif jouit de la *propriété d'autocorrection :* une erreur de calcul isolée ne perturbe pas le résultat final, une approximation erronée étant toujours considérée comme un nouveau vecteur initial. On peut apporter quelques améliorations à cette méthode de manière à accélérer la convergence du processus : tel est l'objet par exemple de la *méthode de Seidel*, qui consiste essentiellement à tenir compte, lors de la k-ième approximation de l'inconnue x_i, des k-ièmes approximations des inconnues $x_1, \ldots, x_{i-1}$ déjà calculées. Il existe d'autres méthodes (*méthode de relaxation* par exemple) qui sont exposées dans la littérature spécialisée.

Résolution approchée des systèmes d'équations non linéaires

Position du problème

Le cas linéaire que nous venons de traiter recouvre un grand nombre de problèmes. Il n'en reste pas moins que certains phénomènes physiques sont régis par des relations non linéaires. D'où l'importance que l'on est en droit d'attacher aux méthodes de résolution de tels systèmes.

Soit donc le système de n équations à n inconnues :

$$\begin{cases} f_1\ (x_1, \ldots, x_n) = 0 \\ f_2\ (x_1, \ldots, x_n) = 0 \\ \quad \vdots \\ f_n\ (x_1, \ldots, x_n) = 0 \end{cases}$$

On peut représenter le système sous la forme abrégée usuelle : $\overrightarrow{f\ (\vec{x})} = \vec{0}$ où f désigne l'application de $\mathbb{R}^n \to \mathbb{R}^n$ définie à l'aide des applications f_i :

$$f\ (\circ) = \begin{bmatrix} f_1\ (\circ) \\ f_2\ (\circ) \\ f_n\ (\circ) \end{bmatrix}$$

Les méthodes qui existent sont assorties de conditions particulièrement contraignantes sur la fonction f. C'est dire que ce genre de méthodes ne répondra qu'à une classe assez limitée de problèmes. Nous présenterons deux algorithmes de résolution : la *méthode de Newton* qui fait appel au calcul différentiel et la *méthode des approximations successives* qui généralise celle que l'on a rencontrée dans le cas linéaire.

La méthode de Newton

Cette méthode s'inspire aussi des principes de la méthode des approximations successives : elle définit un algorithme qui à chaque itération donne une valeur $\vec{x}^{(p)}$ approchée de la solution, qui sert également au calcul de l'approximation suivante $\vec{x}^{(p+1)}$. Cherchons à établir

la règle de calcul. On peut écrire $\vec{x}^{(p+1)} = \vec{x}^{(p)} + \vec{\varepsilon}^{(p)}$. Nous allons chercher à déterminer $\varepsilon^{(p)}$ de telle façon que $x^{(p+1)}$ vérifie « le mieux possible » la condition $f(x) = \vec{0}$, ce qui s'écrit

(7) $$f(\vec{x}^{(p+1)}) = f(\vec{x}^{(p)} + \vec{\varepsilon}^{(p)}) = \vec{0}$$

On suppose que la fonction f est continûment dérivable dans un certain domaine Ω convexe et ouvert de $\mathbb{R}^n$. On peut donc définir la matrice jacobienne de l'application f :

$$W(x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\vec{x}), \ldots, \frac{\partial f_1}{\partial x_n}(\vec{x}) \\ \frac{\partial f_n}{\partial x_1}(\vec{x}) \quad \frac{\partial f_n}{\partial x_n}(\vec{x}) \end{bmatrix}$$

Transformons **(7)** en écrivant le développement limité (vectoriel !) de la fonction f au voisinage du point $\vec{x}^{(p)}$:

$$f(\vec{x}^{(p)} + \varepsilon^{(p)}) = f(\vec{x}^{(p)}) + W(\vec{x}^{(p)}) \cdot \vec{\varepsilon}^{(p)} + \mathcal{O}(\|\vec{\varepsilon}^{(p)}\|)$$

En nous bornant ainsi au premier terme du développement limité, nous pouvons écrire l'égalité suivante qui constitue une approximation de la condition **(7)** :

$$f[\vec{x}^{(p)}] + W(\vec{x}^{(p)}) \cdot \vec{\varepsilon}^{(p)} = \vec{0}$$

en supposant que la matrice $W(\vec{x}^{(p)})$ est inversible (lorsque $\vec{x}^{(p)}$ n'est pas un point singulier de la fonction f), on obtient : $\vec{\varepsilon}^{(p)} = -W(\vec{x}^{(p)})^{-1} \cdot f(\vec{x}^{(p)})$; par conséquent on obtient le vecteur $\vec{x}^{(p+1)}$ selon la règle :

(8) $$\vec{x}^{(p+1)} = \vec{x}^{(p)} - W^{-1}(\vec{x}^{(p)}) \cdot f[\vec{x}^{(p)}]$$

Exemple : soit le système :

$$\begin{cases} x_1^2 + x_2^2 + x_3^2 = 1 \\ 2x_1^2 + x_2^2 - 4x_3 = 0 \\ 3x_1^2 - 4x_2 + x_3^2 = 0 \end{cases}$$

On part du vecteur initial $x_1^{(0)} = x_2^{(0)} = x_3^{(0)} = 0{,}5$, d'où l'on obtient $f(\vec{x}^{(0)}) = \begin{bmatrix} -0{,}25 \\ -1{,}25 \\ -1{,}00 \end{bmatrix}$.

La matrice jacobienne $W(\vec{x})$ est donnée par :

$$W(\vec{x}) = \begin{bmatrix} 2x_1 & 2x_2 & 2x_3 \\ 4x_1 & 2x_2 & -4 \\ 6x_1 & -4 & 2x_3 \end{bmatrix}$$

On a $$W(\vec{x}^{(0)}) = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 1 & -4 \\ 3 & -4 & 1 \end{pmatrix}$$

et $$W^{-1}(\vec{x}^{(0)}) = \begin{pmatrix} \frac{3}{8} & \frac{1}{8} & \frac{1}{8} \\ \frac{7}{20} & \frac{1}{20} & -\frac{3}{20} \\ \frac{11}{20} & -\frac{7}{40} & \frac{1}{40} \end{pmatrix}$$

On calcule ainsi $\vec{x}^{(1)}$ par la forme **(8)**. Le résultat des trois premières itérations est donné dans ce tableau :

(k)	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$
0	0,5	0,5	0,5
1	0,875	0,5	0,375
2	0,789 81	0,496 62	0,369 93
3	0,785 21	0,496 62	0,369 92

Il va sans dire que la convergence de l'algorithme n'est pas toujours assurée et que certaines conditions portant sur l'application f et sur la valeur initiale $\vec{x}^{(0)}$ doivent être réalisées. Ainsi que nous l'avons déjà vu en introduction, la limite de l'algorithme peut dépendre de la valeur initiale.

Il s'agit donc également d'étudier la *stabilité* de la solution par rapport à $\vec{x}^{(0)}$.

La méthode des approximations successives

Soit à résoudre un système non linéaire de la forme

$$\begin{cases} x_1 = \varphi_1(x_1, \ldots, x_n) \\ \vdots \\ x_n = \varphi_n(x_1, \ldots, x_n) \end{cases}$$

que l'on peut écrire sous forme abrégée

(9) $$\vec{x} = \overrightarrow{\varphi(\vec{x})}$$

la solution $\vec{x}^*$ de cette équation, si elle existe, est un *point fixe* de l'application φ. On cherche à savoir à quelles conditions on peut obtenir x^* comme limite du processus itératif défini par :

(10) $$\vec{x}^{(p+1)} = \varphi(\vec{x}^{(p)})$$

Définition : l'application φ de $A \subset \mathbb{R}^n \to \mathbb{R}^n$ sera dite contractante dans A s'il existe une constante a : $0 < a < 1$ telle que

$$\|\varphi(x_1) - \varphi(x_2)\| \leqslant a \|x_1 - x_2\| \quad \forall x_1, x_2 \in A$$

où $\| \ \|$ désigne la norme euclidienne usuelle de $\mathbb{R}^n$. Si l'application φ est contractante dans A, partie fermée de $\mathbb{R}^n$, alors le processus itératif défini en **(10)** converge vers un vecteur $\vec{x}^*$ solution de **(9)**. En outre, cette convergence est indépendante du choix de la valeur initiale $\vec{x}^{(0)}$. Enfin l'erreur absolue commise en stoppant la procédure à l'étape p est majorée selon la formule :

$$\|\vec{x}^* - \vec{x}^{(p)}\| \leqslant \frac{a^p}{1-a} \|x^{(1)} - x^{(0)}\|.$$

La solution $\vec{x}^*$ est alors unique. La méthode des approximations successives peut être également appliquée au système général **(11)** $f(\vec{x}) = \vec{0}$. Considérons le système

(12) $$\vec{x} + f(\vec{x}) = \vec{x}.$$

Toute solution de **(11)** est solution de **(12)**. On peut donc espérer dans certains cas [c'est-à-dire au voisinage de la solution x^* de **(11)**] appliquer cette méthode au cas général en posant $\overrightarrow{\varphi(\vec{x})} = \vec{x} + \overrightarrow{f(\vec{x})}$...

Optimisation

Position du problème

Un problème d'optimisation consiste, d'une façon générale, à déterminer la valeur x^* d'un vecteur $\vec{x}$ de $\mathbb{R}^n$ maximisant — ou minimisant — une fonction *objectif* $f(\vec{x})$. D'autre part le vecteur $\vec{x}^*$ est assujetti à appartenir à une certaine partie K de $\mathbb{R}^n$: $f(\vec{x}) = \underset{x \in k}{\text{Opt}} f(\vec{x})$

où $$f : \mathbb{R}^n \to \mathbb{R}$$

Suivant la nature de l'ensemble K on aura affaire à des types de problèmes différents. On distinguera trois cas :

— $K = \mathbb{R}^n$; il s'agit alors d'un problème d'optimisation sans contraintes ou de recherche d'un *optimum libre* ;

— K est défini par des contraintes de la forme :

$$K = \{\vec{x} \in \mathbb{R}^n) \quad g_j(\vec{x}) = b_j \quad j = 1, \ldots, p\} ;$$

il s'agit alors d'un problème d'optimisation *sans contraintes d'égalité ;*

— K est défini par des contraintes de la forme :

$$K = \{\vec{x} \in \mathbb{R}^n / g_j(\vec{x}) \leqslant b_j \quad j = 1, \ldots, p\}$$

il s'agit dans ce cas d'un problème d'optimisation *sous contraintes d'inégalité.*

Notons au préalable qu'il est toujours possible d'exprimer un problème de maximisation sous la forme d'un problème de minimisation — et réciproquement. En effet, on peut toujours écrire :

$$\text{Max} f(\vec{x}) = -\text{Min}[-f(\vec{x})]$$

Nous restreindrons donc notre exposé au problème de la maximisation. La théorie mathématique va nous permettre de connaître les conditions portant sur la fonction f et sur l'ensemble K qui assurent l'existence (et l'unicité dans certains cas) d'une solution au problème :

(13) $$\underset{\vec{x} \in K}{\operatorname{Max}} f(\vec{x})$$

Théorème :

— *Si f est une fonction continue et si l'ensemble* K *est une partie compacte de* $\mathbb{R}^n$ *(c'est-à-dire fermée et bornée), alors il existe au moins une valeur* $\vec{x}^*$ *de* $\vec{x}$ *solution de* **(13).**

— *Si de plus la fonction f est strictement concave et si l'ensemble* K *est convexe, alors la solution* $\vec{x}^*$ *est unique.*

La première assertion n'est rien d'autre que l'expression du théorème de Weierstrass (voir *Topologie*).

Démontrons la seconde : supposons que $\vec{x}^*$ et $\vec{y}^*$ soient deux solutions distinctes de **(13).** Soit $\vec{z}$ un point du segment joignant $\vec{x}^*$ à $\vec{y}^*$; $\vec{z}$ peut se mettre sous la forme

$$z = \lambda\vec{x}^* + (1 - \lambda)\vec{y}^* \quad \text{où} \quad \lambda \in]0,1[$$

f étant strictement concave, on écrit :

$$f(\vec{z}) = f(\lambda\vec{x}^* + (1-\lambda)\vec{y}^*) > \lambda f(\vec{x}^*) + (1-\lambda) f(\vec{y}^*)$$

or $f(\vec{x}^*) = f(\vec{y}^*)$ par conséquent $f(\vec{z}) > f(\vec{x}^*)$ et x^* ne serait plus solution de **(13).** Pour résoudre numériquement le problème **(13),** il existe deux classes de méthodes que l'on retrouvera dans les 3 cas que nous venons de distinguer : d'une part les *méthodes directes* qui utilisent des conditions *nécessaires* d'optimalité qui s'obtiennent en employant certaines caractérisations issues du calcul différentiel ; d'autre part, les méthodes itératives qui consistent, partant d'un point $\vec{x}_0$ réalisable (c'est-à-dire appartenant à K), à définir une suite de points de K $\vec{x}_1, \ldots, \vec{x}_\nu$ tels qu'à chaque étape on « améliore » la fonction objectif : $\varphi(\vec{x}_0) < \varphi(\vec{x}_1) \ldots < \varphi(\vec{x}_\nu) < \ldots$ Si la procédure converge, on est en droit d'espérer que la limite soit solution du problème **(13).** Encore faudra-t-il s'assurer que l'on n'obtient pas ainsi un optimum local qui dépendrait de la valeur initiale. Les méthodes itératives diffèrent entre elles par la procédure employée pour passer d'un point $\vec{x}_\nu$ de l'algorithme au suivant $\vec{x}_{\nu+1}$. En général, on cherchera $\vec{x}_{\nu+1}$ sous la forme $\vec{x}_{\nu+1} = \vec{x}_\nu + t_\nu \cdot \vec{d}_\nu$ où $t_\nu \in \mathbb{R}$ et $\vec{d}_\nu$ est un vecteur de $\mathbb{R}^n$ qui représente une *direction admissible* choisie de telle façon que, en la suivant, on augmente la valeur de la fonction f ; t_ν représente le « pas » du déplacement que l'on détermine de manière à améliorer réellement la fonction f (une trop grande valeur de t_ν pourrait conduire à une décroissance de f) *tout en restant dans le domaine admissible K.* Le tableau ci-dessous résume les méthodes les plus couramment utilisées dans les problèmes d'optimisation.

Nous n'évoquerons ici qu'une partie des méthodes de ce tableau. En particulier, nous n'aborderons pas ici la programmation linéaire (et la programmation quadratique) qui fera l'objet de développements ultérieurs (voir *Mathématiques et science économique*).

Optimisation sans contraintes - Méthode directe

Soit donc le programme suivant :

(14) $$\underset{x \in \mathbb{R}^n}{\operatorname{Max}} f(x).$$

Nous supposons que la fonction objectif f est de classe C^2.

Montrons qu'une *condition nécessaire* pour que $\vec{x}^*$ soit un maximum de f s'écrit

(15) $$\frac{\partial f}{\partial x_i}(\vec{x}^*) = 0 \quad i = 1, \ldots, n.$$

Cette condition n'est en effet nullement suffisante puisqu'elle serait également vérifiée pour un minimum ou pour ce que l'on appelle un *point-selle*. En outre elle est une condition nécessaire d'*optimalité locale*.

Au voisinage de x^*, on écrit le développement de Taylor de la fonction f à l'ordre 2 :

(16) $$f(\vec{x}^* + \vec{h}) - f(\vec{x}^*) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\vec{x}^*) \cdot h_i + \mathcal{O}(\|\vec{h}\|).$$

En appelant *vecteur gradient* le vecteur $\nabla f(\vec{x}^*)$ de composantes $\left[\frac{\partial f}{\partial x_1}(\vec{x}^*), \frac{\partial f}{\partial x_2}(\vec{x}^*), \ldots, \frac{\partial f}{\partial x_n}(\vec{x}^*)\right]$, on réécrit l'égalité **(16).**

(17) $$f(\vec{x}^* + \vec{h}) - f(\vec{x}^*) = \nabla f(\vec{x}^*) \cdot \vec{h} + \mathcal{O}(\|\vec{h}\|),$$

où « · » désigne le produit scalaire dans $\mathbb{R}^n$. Dire que $\vec{x}^*$ est un maximum local, c'est dire qu'il existe un voisinage V $(\vec{0})$ tel que :

$$f(\vec{x}^* + \vec{h}) \leqslant f(\vec{x}^*) \quad \forall \vec{h} \in V(\vec{0}).$$

Pour $\|\vec{h}\|$ suffisamment petit (c'est-à-dire inférieur à ε fixé à l'avance) il vient :

(18) $$\nabla f(\vec{x}^*) \cdot \vec{h} \leqslant 0 \quad \forall h \quad \|\vec{h}\| \leqslant \varepsilon.$$

Il est clair que cette inégalité entraîne $\overrightarrow{\nabla f(x^*)} = \vec{0}$. En effet, il suffit d'écrire l'inégalité **(18)** successivement pour $h = (0, \ldots h_i, \ldots 0)$ et $(0, \ldots 0, -h_i, \ldots 0)$ pour i variant de 1 à n.

$$f(\vec{x}^*) = \underset{\vec{x} \in \mathbb{R}^n}{\operatorname{Max}} f(\vec{x}) \Rightarrow \nabla f(\vec{x}^*) = \vec{0}.$$

La recherche du maximum peut donc se ramener à la résolution du système **(15).** A cet égard, pour résoudre le système, on peut développer une de ces méthodes présentées dans le chapitre précédent. Comme nous l'avons dit, rien n'indique que la résolution d'un tel système conduise à une solution du problème de maximisation. Les deux problèmes ne sont pas équivalents. La seule chose que l'on puisse affirmer, c'est que l'ensemble des solutions **(14)** est inclus dans celui des solutions de **(15).**

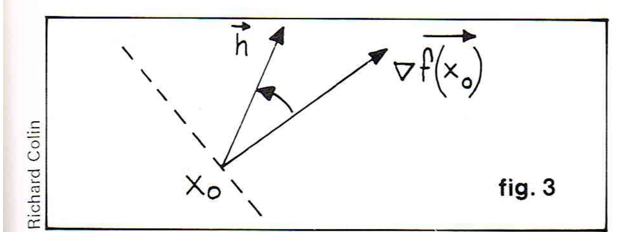
Optimisation sans contraintes - Méthode du gradient

La méthode du gradient est une méthode itérative qui est fondée sur les considérations suivantes : soit un point x_0 de $\mathbb{R}^n$. Certains déplacements $\vec{h}$ au voisinage de ce point vont produire un accroissement de la fonction f. Nous pouvons caractériser ces déplacements en écrivant la formule **(17)** :

$$f(x_0 + h) - f(x_0) = \overrightarrow{\nabla f(x_0)} \cdot \vec{h} + \mathcal{O}(\|h\|)$$

Contraintes	Méthodes directes	Méthodes itératives
Sans contraintes	**Annulation des dérivées partielles**	**Méthode du gradient** **Méthode du gradient conjugué**
Contraintes d'égalité	**Multiplicateurs de Lagrange**	**<u>Contraintes linéaires</u>** **Programmation linéaire (simplex)** **Programmation, quadratique (Danzig, Wolfe)** **Méthode du gradient projeté (Rosen)**
Contraintes d'inégalité	**Multiplicateurs de Kuhn et Tücker**	**<u>Contraintes non linéaires</u>** **Méthode des directions admissibles** **Méthodes des pénalités**

▶ ***Tableau récapitulatif des méthodes les plus couramment utilisées dans les problèmes d'optimisation.***



◀ **Figure 3 : la fonction f augmentera pour tout déplacement $\vec{h}$ suffisamment petit vérifiant : $\overrightarrow{\nabla f(x_0)} \cdot \vec{h} \geqslant 0$.**

Au premier ordre, l'on peut dire que la fonction f augmentera pour tout déplacement $\vec{h}$ suffisamment petit vérifiant : $\overrightarrow{\nabla f(x_0)} \cdot \vec{h} \geqslant 0$, c'est-à-dire pour tout vecteur $\vec{h}$ faisant un angle aigu avec le vecteur gradient (*fig.* 3) : le déplacement qui assure le plus grand accroissement de la fonction est celui qui correspond à la direction du gradient lui-même.

La méthode du gradient consiste donc, à chaque itération, à suivre la direction du gradient aussi longtemps que la fonction augmente : plus précisément, à l'itération ν on se trouve en un point $\vec{x}_\nu$. On cherche $\vec{x}_{\nu+1}$ sous la forme :

$$\vec{x}_{\nu+1} = \vec{x}_\nu + t_\nu \ \overrightarrow{\nabla f(\vec{x}_\nu)} \quad t_\nu \in \mathbb{R}$$

t_ν est la valeur du pas qui donne la plus grande valeur de la fonction f selon la direction $\overrightarrow{\text{grad} f(x_\nu)}$. Autrement dit, t_ν est solution du problème de maximisation à une variable :

$$\underset{t \in \mathbb{R}}{\text{Max}} \, f(\vec{x}_\nu + t \nabla f(\vec{x}_\nu))$$

Exemple : soit à résoudre le problème de maximisation à 2 variables x_1 et x_2

$$\underset{(x_1, x_2) \in \mathbb{R}^2}{\text{Max}} \quad f(x_1, x_2) = -(x_1 - 2)^2 - 2(x_2 - 1)^2$$

La solution est bien sûr évidente : $x^* = 2$ et $x_2^* = 1$. En tout point x les composantes du vecteur gradient sont données par

$$\nabla f(x) = \begin{cases} -2(x_1 - 2) \\ -4(x_2 - 1) \end{cases}$$

Partons d'une valeur initiale $\vec{x}^0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$; $\nabla f(\vec{x}_0) = \begin{pmatrix} 4 \\ 4 \end{pmatrix}$.

On détermine t_0 qui maximise $f\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix} + t\begin{pmatrix} 4 \\ 4 \end{pmatrix}\right]$, soit :

$$\frac{d}{dt}[(4t - 2)^2 + 2(4t - 1)^2] = 0, \quad \text{ce qui donne } t_0 = \frac{1}{3}.$$

On a donc $\vec{x}^1 = \begin{pmatrix} \frac{4}{3} \\ \frac{4}{3} \end{pmatrix}$. On calcule ici $\nabla f(x^1) = \begin{pmatrix} \frac{4}{3} \\ -\frac{4}{3} \end{pmatrix}$

puis t_1, etc.

Les calculs correspondant aux trois premières itérations sont rassemblés dans le tableau suivant :

	x_ν	$\nabla f(x_\nu)$	t_ν
0	(0,0)	(4,4)	1/3
1	(4/3, 4/3)	(4/3, - 4/3)	1/3
2	(16/9, 8/9)	(4/9, 4/9)	1/3
3	(52/27 , 28/27)	....	....

Cette méthode comporte les avantages et les inconvénients de toutes les méthodes itératives tels que nous les avons évoqués dans les chapitres précédents. L'exemple que nous venons de développer ne pose quant à lui aucun problème d'unicité puisque la fonction objectif choisie est une fonction concave.

Optimisation sous contraintes d'égalité - Multiplicateurs de Lagrange

Soit à résoudre le problème d'optimisation suivant :

$$\text{Max} \, f(\vec{x}) \qquad n \geqslant m$$
$$|g_j(\vec{x}) = b_j \qquad j = 1, \ldots m$$

que l'on peut réécrire sous une forme vectorielle :

$$\text{Max} \, f(\vec{x})$$
$$|\vec{g}(\vec{x}) = \vec{b}$$

où g représente une application de $\mathbb{R}^n \to \mathbb{R}^m$ et $\vec{b}$ un vecteur de $\mathbb{R}^m$. L'idée qui vient immédiatement à l'esprit consiste à tenter de se ramener à un problème sans contraintes en exprimant certaines variables en fonction de $n - m$ variables libres à l'aide des m contraintes d'égalité. Cette démarche est directement fondée sur le *théorème des fonctions implicites;* l'application de ce théorème en un point $\vec{x}^*$, maximum local, permet (sous les bonnes hypothèses) de trouver des considérations nécessaires d'optimalité qui généralisent donc celles que nous avons rencontrées dans le paragraphe consacré à la méthode directe : le développement de cette méthode aboutit à définir des coefficients réels $\lambda_1, \ldots, \lambda_m$ (ou un vecteur $\vec{\lambda}$ de $\mathbb{R}^m$) pour chaque contrainte, que l'on appelle *multiplicateurs de Lagrange*, et à introduire une fonction nouvelle : $L(x_1, \ldots, x_n, \lambda_1, \ldots, \lambda_m)$ notée $L(\vec{x}, \vec{\lambda})$,

définie par : $L(\vec{x}, \vec{\lambda}) = f(\vec{x}) - \sum_{j=1}^{m} \lambda_j [g_j(\vec{x}) - \vec{b}_j]$

ou mieux : $L(\vec{x}, \vec{\lambda}) = f(\vec{x}) - \vec{\lambda} \cdot (g(\vec{x}) - \vec{b})$.

La condition nécessaire d'optimalité s'exprime en écrivant que $\vec{x}^*$ doit satisfaire les contraintes :

(19) $$g(\vec{x}^*) - b = 0$$

et que le gradient (par rapport à $\vec{x}$) du *lagrangien* est nul :

(20) $$\nabla f(x^*) - \vec{\lambda}^* \cdot \nabla g(x^*) = \vec{0}.$$

On peut regrouper les conditions **(19)** et **(20)** sous une forme plus condensée :

(21) $$\begin{cases} \nabla_x L(\vec{x}^*, \vec{\lambda}^*) = \vec{0}. \\ \nabla_\lambda L(\vec{x}^*, \vec{\lambda}^*) = \vec{0}. \end{cases}$$

Il s'agit d'un système de $n + m$ équations à $n + m$ inconnues $x_1, \ldots, x_n$ et $\lambda_1, \ldots, \lambda_m$. On notera la symétrie que ces formules accordent à $\vec{x}$ et à $\vec{\lambda}$; celle-ci traduit le fait qu'un problème d'optimisation peut se formuler dans le cadre plus général de la dualité. Nous retrouverons ce point de vue dans le chapitre *Mathématiques et science économique*.

La méthode des multiplicateurs de Lagrange, qui, répétons-le, ne fournit qu'une condition nécessaire d'optimalité, a été généralisée dans le cas de contraintes d'inégalité, c'est-à-dire pour résoudre un problème de la forme :

$$\text{Max} \, f(\vec{x})$$
$$g_j(\vec{x}) \leqslant b_j \quad j = 1, \ldots, m$$
$$\vec{x}_i \geqslant 0 \qquad i = 1, \ldots, n$$

Notons la contrainte de positivité des variables. On définit alors, de manière analogue, une série de m coefficients appelés $\mu_1, \ldots, \mu_m$, appelés *multiplicateurs de Kühn et Tücker*, et qui permettent de définir un ensemble de conditions qui généralisent **(21)**.

Maximisation sous contraintes d'inégalité - Méthodes itératives

Méthode Rosen (gradient projeté)

La méthode du gradient projeté s'applique au cas où les *contraintes* sont *linéaires :*

$$\text{Max} \ f(\vec{x})$$
$$\vec{A}x \leqslant \vec{b}$$
$$\vec{x} \geqslant 0$$

A représente une matrice (m, n) et $\vec{b}$ un vecteur de $\mathbb{R}^n$. La méthode s'inspire des principes de la méthode du gradient adaptée à la recherche d'un optimum libre : partant d'un point réalisable $\vec{x}^0$, on « améliore » la fonction objectif en se déplaçant suivant la direction

$$\vec{d}_0 = \nabla f\,(\vec{x}_0).$$

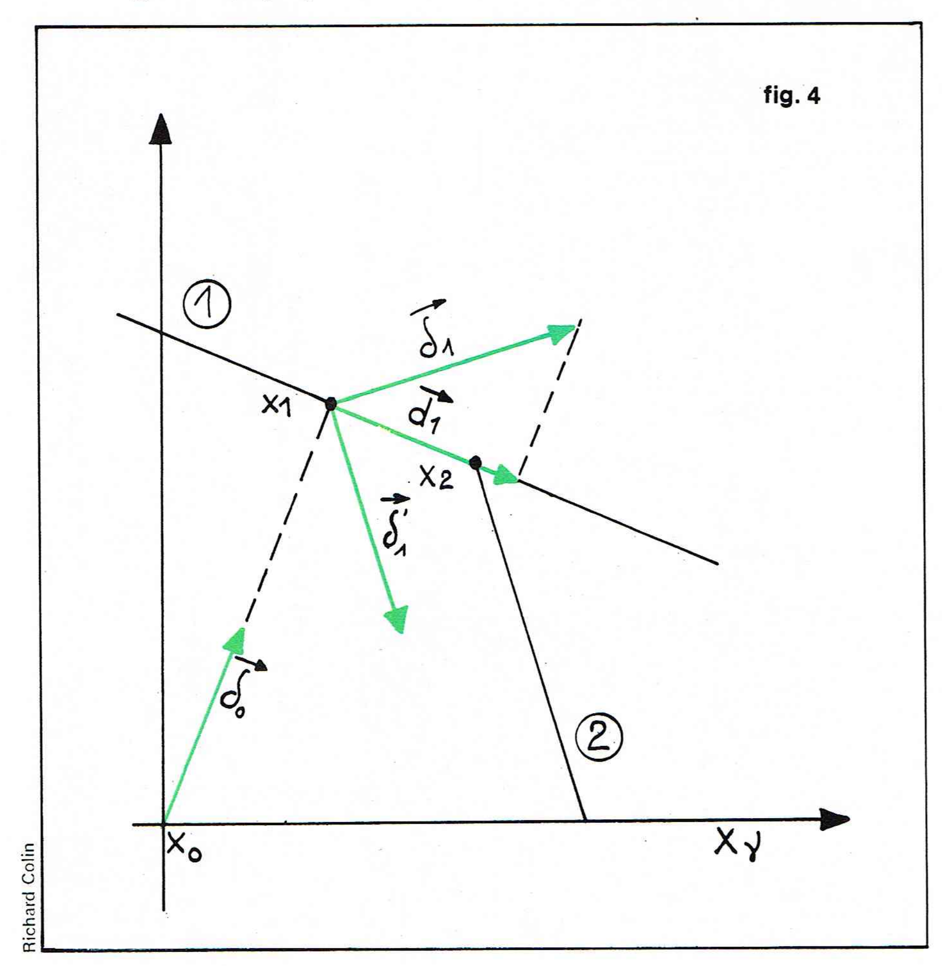
On se déplace tant que la fonction objectif croît et tant que l'on n'atteint pas une contrainte (la *contrainte* sera dite alors *saturée*, mais il peut y en avoir plusieurs). Si l'on atteint une contrainte en un point $\vec{x}_1$, on calcule le vecteur $\nabla f\,(\vec{x}_1)$ que *l'on projette sur la ou les contraintes saturées*. La projection $\vec{d}_1$ ainsi déterminée fournit la direction du déplacement de l'itération suivante. En fait, une fois calculé le vecteur gradient $\nabla f\,(\vec{x}_0) = \vec{\delta}_\nu$, on considère uniquement sa projection sur l'ensemble des contraintes saturées en $\vec{x}_\nu$ qui « empêcherait » de prendre la direction $\vec{\delta}_\nu$: ainsi, sur la *figure* 4, en $\vec{x}_1$, si le gradient de *f* est égal à $\vec{\delta}'_1$, on ne projettera pas sur la contrainte 1, mais on gardera le déplacement $\vec{\delta}'_1$.

Certains artifices de calcul matriciel permettent, à chaque itération, de calculer explicitement le vecteur projection, c'est-à-dire la direction $\vec{d}_\nu$ à suivre pour atteindre $\vec{x}_{\nu+1}$. La linéarité des contraintes assure la convexité de l'ensemble des points admissibles.

D'autres méthodes sont couramment pratiquées pour la résolution de problèmes d'optimisation. Elles s'inspirent toutes des mêmes arguments que celle du gradient projeté. Lorsque les contraintes ne sont plus linéaires (par exemple convexes, de manière à assurer la convexité de l'ensemble des points réalisables), la recherche de directions admissibles $\vec{d}_\nu$ à partir d'un point réalisable $\vec{x}_\nu$ est beaucoup plus délicate : il est alors extrêmement malaisé de se déplacer sur des contraintes non linéaires car cela nécessite une série de calculs d'une rare complexité. Il existe principalement deux types de méthodes pour résoudre ce genre de problèmes : la méthode des directions admissibles, et celle des pénalités.

Méthode des directions admissibles (Zoutendijk)

Si, lors d'une itération, on se trouve sur la frontière du domaine en un point $\vec{x}_\nu$, on déterminera la direction admissible $\vec{d}_\nu$ qui fait revenir à l'intérieur du domaine, c'est-à-dire que l'on ne se déplace plus selon la direction du gradient $\nabla f\,(\vec{x}_\nu)$ mais selon une direction $\vec{d}_\nu$ qui vérifie $\nabla f\,(\vec{x}_\nu) \cdot \vec{d}_\nu > 0$, de telle façon, néanmoins, que l'on reste dans le domaine des points réalisables.



▼ ***Figure 4 :* *voir développement dans le texte*** **Méthode Rosen (gradient projeté).**

Méthode des pénalités

La méthode des pénalités est une méthode assez largement utilisée dans la pratique. La formulation la plus courante semble en être celle de Fiacco-Mac Cormick. Sans entrer dans les détails, donnons le principe de ce genre de méthode. Soit donc un problème d'optimisation du type :

$$\textbf{(22)} \qquad \text{Max } f\,(\vec{x}) \qquad g_i\,(\vec{x}) > 0$$

Les contraintes de positivité des variables sont incluses dans les fonctions $g_i\,(\vec{x})$. La méthode consiste à modifier la fonction objectif d'une certaine manière, de façon à remplacer le problème initial **(22)** par une séquence de problèmes sans contraintes (que l'on peut résoudre par une méthode du type gradient) dont les solutions convergent vers l'optimum recherché. Le principe de pénalisation revient à modifier la fonction objectif en y faisant figurer les fonctions qui définissent les contraintes de telle sorte que l'on soit « fortement pénalisé » si l'on s'aventure en dehors du domaine réalisable

$$P = \{\vec{x}/g_i\,(\vec{x}) \geqslant 0\}.$$

Considérons par exemple la fonction de pénalité définie par :

$$\Phi\,(y) = \begin{cases} 0 & y \geqslant 0 \\ -\infty & y < 0 \end{cases}$$

On modifie alors la fonction objectif de la manière suivante :

$$\Psi\,(x) = f\,(x) + \sum_{j=1}^{m} \Phi\,[g_j\,(\vec{x})]$$

Si l'on cherche à minimiser *sans contraintes* la fonction Ψ, on a de fortes chances de rester dans le domaine P : si l'un des $g_j\,(\vec{x})$ devient < 0, alors $\Phi\,[g_j\,(\vec{x})]$ devient $-\infty$ et la violation d'une contrainte est de ce fait fortement pénalisée. Il serait prématuré de crier victoire et de croire que l'on peut ainsi remplacer tout problème d'optimisation *avec contraintes* par un problème *sans contraintes*. En effet la fonction Ψ n'est pas continue puisque la fonction Φ ne l'est point. Impossible donc d'utiliser une méthode de gradient, car, dès que l'on atteint une contrainte, on ne peut plus calculer le gradient. On est donc tenu de choisir une fonction de pénalisation continue; par exemple :

$$\Phi\,(y) = \begin{cases} 0 & y \geqslant 0 \\ -y^2 & y < 0 \end{cases}$$

Mais dans ce cas, on peut s'attendre à ce que certaines contraintes soient violées car la « pénalisation » n'est pas suffisamment forte. Pour éviter donc le plus grand nombre de ces violations, on pondère largement par un coefficient $K > 0$ en posant :

$$\textbf{(23)} \quad \text{Max}\left[f\,(\vec{x}) + K \sum_{j=i}^{m} \Phi\,[g_j\,(\vec{x})]\right] = \underset{\vec{x}}{\text{Max}}\,\Psi\,(\vec{x}, K).$$

La méthode consiste à résoudre le problème **(23)** pour différentes valeurs de K : on considère une suite décroissante : $K_0 > K_1 > \ldots > K_\nu > K_{\nu+1}$;

soit $\vec{x}^*\,(K_0)$, $\vec{x}^*\,(K_1)$, ..., $x^*\,(K_\nu)$ les solutions correspondantes du problème **(23)**. La présence de la pénalisation nous assure que tous ces points sont intérieurs au domaine P. Il reste à espérer que la suite $\vec{x}^*\,(K_\nu)$ converge vers la solution $\vec{x}^*$ du problème initial **(22)**.

$$\lim_{\nu \to \infty} x^*\,(K_\nu) = x^*\,?$$

$$\lim_{\nu \to \infty} \Psi\,[x^*\,(K_\nu)] = f\,(\vec{x}^*)\,?$$

Cette convergence est réalisée si, entre autres, les fonctions *f*, g_j et Φ vérifient certaines hypothèses de convexité et de différentiabilité.

Approximation - Interpolation

Les fonctions les plus connues et les plus utilisées dans les calculs sont celles qui sont données de manière analytique et qui peuvent être calculées par des ordinateurs à l'aide des opérations élémentaires préprogrammées.

Ainsi en est-il de la fonction polynôme (de degré n) :

(24) $\pi_n(x) = a_0 + a_1x + a_2x^2 + \dots a_nx^n$

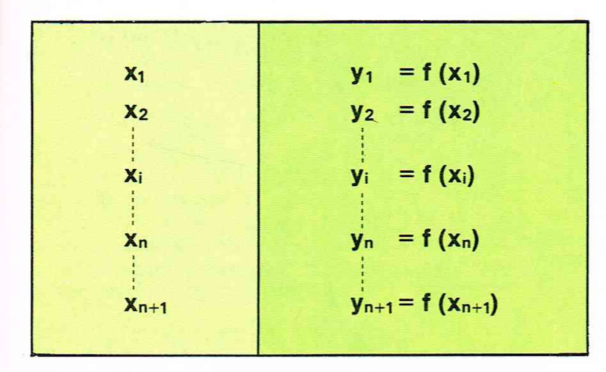
des fonctions rationnelles de la forme :

(25) $$R^n(x) = \frac{a_0 + a_1x + \dots a_nx^n}{b_0 + b_1x + \dots b_nx^n}$$

ou encore des polynômes trigonométriques :

(25') $$T_n(x) = \frac{1}{2}a_0 + \sum_{k=1}^{n}(a_k \cos kx + b_k \sin kx).$$

Il arrive que, dans la pratique, l'on se trouve en présence de fonctions que l'on ne peut exprimer sous forme analytique de manière simple et opérationnelle. Il arrive aussi que l'on ne connaisse la fonction que par les valeurs qu'elle prend en un nombre déterminé de points $x_1, \dots, x_n$. Ces valeurs sont données dans un tableau **(26)** du type :



x_1	$y_1 = f(x_1)$
x_2	$y_2 = f(x_2)$
⋮	⋮
x_i	$y_i = f(x_i)$
⋮	⋮
x_n	$y_n = f(x_n)$
⋮	⋮
x_{n+1}	$y_{n+1} = f(x_{n+1})$

Si l'on souhaite calculer la valeur prise par la fonction en un point $x \neq x_i$, il faudra se contenter d'une valeur approchée. Pour cela on cherchera à remplacer la fonction f par une fonction connue aisément calculable, un polynôme $\pi_m(x)$. On pourrait tout aussi bien chercher des fonctions du type **(25)** ou **(25')**. Le tout est de choisir un polynôme $\pi_m(x)$ de telle sorte qu'il ne soit pas trop « différent » (dans un sens à préciser) de la fonction f. A cet égard, on peut envisager deux types de méthodes : l'approximation et l'interpolation.

L'approximation

Nous avons eu déjà l'occasion de rencontrer (voir *Analyse des données*) un exemple d'approximation : la *régression linéaire*. Plus généralement, on peut chercher à « approcher » les points (x_i, y_i) par une courbe du 2e, 3e ou m-ième degré (*fig.* 5). Il s'agira donc de déterminer le polynôme $\pi_m^*(x) = a_0^* + a_1^*x + \dots a_m^*x^m$ qui minimise la distance euclidienne $\sum_{i=1}^{n}[y_i - \pi_m(x_i)]^2$.

Appelons $\Phi(a_1, a_2, \dots, a_m)$ cette expression ; le problème revient donc à résoudre le problème d'optimisation :

$$\underset{a_1, \dots a_m \in \mathbb{R}^m}{\text{Min}} \Phi(a_1, a_2, \dots, a_m).$$

On se trouve donc en présence d'un problème sans contraintes que l'on résoudra en usant des méthodes décrites dans le chapitre précédent.

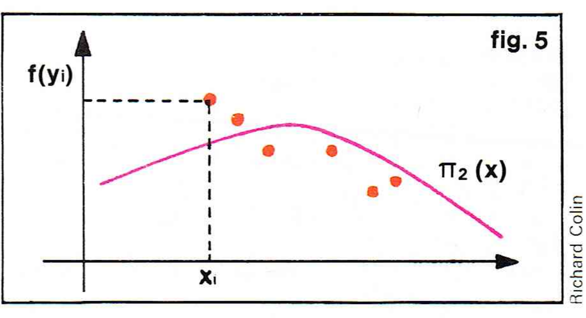
L'interpolation

Étant donné un tableau tel que **(26)**, on cherche à trouver un polynôme $\pi_m(x)$ tel que :

$$\pi_m(x_i) = f(x_i) \qquad \forall i = 1, \dots, n$$

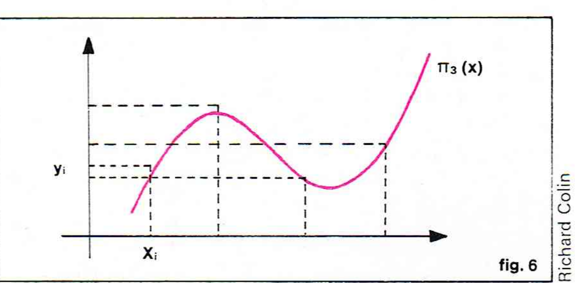
Autrement dit, il s'agit de faire passer la courbe de la fonction $\pi_m(x)$ par tous les points (x_i, y_i) [*fig.* 6] :

Il est possible de démontrer que le polynôme de plus bas degré passant par $n + 1$ points est, au plus, de degré n. Nous allons donner la *formule d'interpolation de Newton*.



Richard Colin

◀ ***Figure 5 ;** l'approximation : **courbe de la fonction** π_2 **(x).***



Richard Colin

◀ ***Figure 6 ;** l'interpolation : **courbe de la fonction** π_3 **(x).***

Pour cela, il nous faut introduire quelques notions liminaires.

Dérivées discrètes d'ordre h

Étant donné deux points du tableau x_i, x_{i+1}, nous définirons la *dérivée discrète du premier ordre* $f[x_i, x_{i+1}]$ par la formule :

$$f[x_i, x_{i+1}] = \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i}$$

Étant donné maintenant 3 points x_{i-1}, x_i, x_{i+1}, nous définirons la *dérivée discrète du deuxième ordre*

$$f[x_{i-1}, x_i, x_{i+1}]$$

▼ ***Le calculateur scientifique du centre de calcul analogique de la faculté des sciences d'Orsay (France).***



Parimage

de la manière suivante :

$$f\,[x_{i-1}, x_i, x_{i+1}] = \frac{f\,[x_{i+1}, x_i] - f\,[x_i, x_{i-1}]}{x_{i+1} - x_i}$$

Plus généralement, étant donné $k + 1$ points

$$(x_i, x_{i+1}, \ldots, x_{i+k+1})$$

on appellera *dérivée discrète d'ordre k* l'expression :

$$f\,[x_i, x_{i+1}, \ldots, x_{i+k+1}] = \frac{f\,[x_{i+1}, \ldots, x_{i+k+1}] - f\,[x_i, \ldots, x_{i+k}]}{x_{i+k+1} - x_i}$$

Formule d'interpolation de Newton

Considérons la dérivée discrète première sur l'intervalle $[x, x_1]$: $f(x) = f(x_1) + (x - x_1)\, f\,[x_1, x]$; on peut remplacer $f\,[x_1, x]$ par $f\,[x_2, x_1] + (x - x_2)\, f\,[x_1, x_2, x]$; en continuant ainsi le procédé, on obtient l'égalité :

$$f(x) = f(x_1) + (x - x_1)\, f\,[x_1, x_2] + (x - x_1)(x - x_2)\, f\,[x_1, x_2, x_3] + \ldots + \ldots (x - x_1) \ldots (x - x_n)\, f\,[x_1, \ldots, x_{n+1}] + \mathrm{E}(x)$$

où $\mathrm{E}(x)$ est le reste qui est égal à

$$(x - x_1) \ldots (x - x_{n+1})\, f\,[x_1, \ldots, x_{n+1}, x].$$

Ainsi donc, on peut écrire la formule d'interpolation de Newton :

$$\pi_n(x) = f(x_1) + (x - x_1)\, f\,[x_1, x_2] + \ldots + (x - x_1) \ldots (x - x_m)\, f\,[x_1, \ldots, x_{n+1}]$$

$\pi_n(x)$ est un polynôme de degré n qui passe par les $n + 1$ points (x_i, y_i). La procédure de calcul du polynôme $\pi_n(x)$ se simplifie notablement lorsque les points x_i sont répartis selon des intervalles réguliers sur $[a, b]$:

$$x_2 - x_1 = x_3 - x_2 = \ldots = x_{n+1} - x_n = k.$$

On peut définir alors la différence finie du premier ordre Δy_i par :

$$\Delta y_i = y_i - y_{i-1} = f(x_i) - f(x_{i-1})\,;$$

et à l'ordre n :

$$\Delta^n y_i = \Delta^{n-1} y_i - \Delta^{n-1} y_{i-1}.$$

La formule de Newton s'écrit alors, en partant du point (x_{n+1}, y_{n+1}) :

$$\pi_n(x) = y_{n+1} + (x - x_{n+1})\frac{\Delta y_{n+1}}{k} +$$

(27) $$(x - x_{n+1})(x - x_n)\frac{\Delta^2 y_{n+1}}{2\,!\,k^2} + \ldots + (x - x_{n+1}) \ldots (x - x_2)\frac{\Delta^n y_{n+1}}{n\,!\,k^n}$$



Erich Hartmann - Magnum

Applications au calcul d'intégrales

Les méthodes d'interpolation et d'approximation que l'on vient d'étudier dans les paragraphes précédents sont d'une grande importance dans le calcul numérique des intégrales définies. Soit donc une fonction $f(x)$ donnée sous forme analytique, dont on ne connaît pas de primitives mais dont on souhaite calculer l'intégrale sur un intervalle $[a, b]$ $(a > -\infty, b < +\infty)$

$$\mathrm{I} = \int_a^b f(x)\, dx.$$

On peut chercher une valeur approchée de I : pour cela, on découpe l'intervalle $[a, b]$ selon n intervalles égaux $[x_i, x_{i+1}]$ tels que $x_{i+1} = x_{i+k}$

$$x_1 = a, \ldots, x_{n+1} = b \qquad k = \frac{b - a}{(n + 1)}$$

La méthode consiste à choisir une bonne interpolation $\tilde{f}$ de la fonction f sur $[a, b]$ et à prendre comme valeur approchée de I la somme :

$$\int_a^b f(x)\, dx \simeq \sum_{i=1}^{n} \int_{x_i}^{x_{i+1}} \tilde{f}(x)\, dx.$$

Par exemple on pourra prendre pour fonction $\tilde{f}$ le polynôme $\pi_n(x)$ donné par la formule **(27).** L'on obtient ainsi :

$$\int_{x_i}^{x_{i+1}} f(x)\, dx = \int_{x_i}^{x_{i+1}} \Big[f(b) + \frac{(x - b)}{k}\Delta f(b) + \frac{(x - b)(x - x_n)}{2\,!\,k^2}\Delta^2 f(b) + \ldots + \frac{\ldots (x - b)(x - x_n)(\ldots)(x - x_2)}{n\,!\,k^n}\Delta^n f(b)\Big] + \int_{x_{i-1}}^{x_i} \mathrm{E}^n(x)\, dx.$$

Si l'on pose $x = b + \alpha k$ $(\alpha < 0)$,

$$\int_{x_i}^{x_{i+1}} f(x)\, dx = \Big[f(b) + \frac{\alpha^2}{2}\Delta f(b) + \frac{\alpha^2}{12}(2\alpha + 3)\,\Delta^2 f(b) + \ldots \Big]_{i-(n+1)}^{i-n} + \frac{k^{n+2}}{(n + 1)\,!}\int_{i-(n+1)}^{i-n} \alpha(\alpha + 1) + \ldots (\alpha + n)\, f^{(n+1)}(\xi)\, d\alpha$$

Cette formule est connue sous le nom de *formule de Newton-Cotes*. On peut l'établir d'une manière plus générale en calculant les coefficients du polynôme en un point quelconque du découpage $(x_j, f(x_j))$. Cette formule permet également de connaître une majoration de l'erreur commise grâce à la donnée du dernier terme.

Intégration des équations différentielles

Position du problème

Un des problèmes les plus couramment traités en calcul numérique est la résolution d'équations différentielles du type :

(28) $$\left[\frac{dx}{dt} = \dot{x}(t) = f\,[x(t), t]\right] \qquad x(t_0) = x_0$$

Les conditions d'existence et d'unicité d'une solution $x^*(t)$ de ce type d'équation ont déjà été présentées et sont connues (voir *Analyse*). Il arrive fréquemment que l'équation différentielle **(28)** ne relève pas d'un type connu (linéaire, de Riccati, etc.). L'on ne possède donc aucune formule analytique permettant d'en obtenir la solution. L'on cherche alors une solution approchée de l'équation ; différentes méthodes de résolution existent ainsi qui, toutes, sont fondées sur le principe suivant : on commence par « discrétiser » l'intervalle $[t_0, t]$ sur lequel on veut intégrer l'équation **(28)** en le subdivisant en N intervalles de longueur $h = \dfrac{t - t_0}{\mathrm{N}}$ au moyen

de N + 1 points intermédiaires $t_0, t_1, \ldots$ avec $t_i = t_{i-1} + h$ et $t_N = t$.

On cherche à attribuer à tous ces points une suite de (N + 1) valeurs $\tilde{x}_0, \tilde{x}_1, \ldots, \tilde{x}_N$ qui constituent des approximations des valeurs prises par la solution $x^*(t)$ de l'équation **(28)**. On obtient ainsi un tableau **(29)** des valeurs approchées :

$$\textbf{(29)} \qquad \begin{matrix} t_0 & \tilde{x}_0 \\ t_1 & \tilde{x}_1 \\ \vdots & \vdots \\ t_N & \tilde{x}_N \end{matrix}$$

Toutes les méthodes consistent à déterminer la valeur $\tilde{x}_{n+1}$ à partir des valeurs $\tilde{x}_0, \ldots, \tilde{x}_n$ grâce à une relation du type :

$$\tilde{x}_{n+1} = F(\tilde{x}_0, \tilde{x}_1, \ldots, \tilde{x}_{n+1};\ n, f).$$

La méthode sera dite implicite si F ne dépend pas de $\tilde{x}_{n+1}$: $\tilde{x}_{n+1} = F(\tilde{x}_0, \tilde{x}_1, \ldots, \tilde{x}_n, n, f)$; elle sera dite explicite dans le cas contraire. Les méthodes diffèrent dans le choix de la fonction F. Cette fonction ne peut être choisie de manière arbitraire : elle doit satisfaire notamment certaines conditions de *compatibilité* avec l'équation **(28)** et de *convergence*.

Conditions de compatibilité

Les valeurs approchées $(\tilde{x}_0, x_1, \tilde{x}_N)$ ne coïncident pas en général avec les valeurs théoriques exactes

$$(x_0^*, \ldots, x_N^*) \qquad [x_n^* = x^*(t_n)]$$

et donc la quantité $l_{n+1} = x_{n+1}^* - F(x_0^*, x_1^*, \ldots, x_n^*)$ est $\neq 0$. Le rapport $\tau_{n+1} = \dfrac{l_{n+1}}{h}$ sera appelé *erreur de troncature locale à la* $(m + 1)$-ième étape. Une méthode sera dite compatible avec l'équation différentielle **(28)** si elle satisfait aux conditions :

(a) $F(\tilde{x}_0, \tilde{x}_1, \ldots, \tilde{x}_{n+1};\ n, 0) = 0$

(b) $|F(\tilde{x}_0, \ldots, \tilde{x}_{nM};\ n, f) - F(\tilde{x}'_0, \ldots, \tilde{x}'_{n+1};\ n, f)| \leqslant C \sum_{i=0}^{n+1} |\tilde{x}_i - \tilde{x}'_i|\ ;\ C > 0$

(c) $\lim_{h \to 0} \tau = 0$ où $\tau = \underset{1 \leqslant i \leqslant N+1}{\text{Max}} |\tau_i|$

La qualité τ mesure *l'erreur de troncature de la méthode ;* elle mesure la qualité de l'approximation choisie. Il est certain que, dans le cas général, ni τ_{n+1}, ni l_{n+1} ne sont donnés sous forme analytique. L'on ne peut en connaître que des majorations qui permettent cependant d'établir la condition **(c)**.

Conditions de convergence

Soit une approximation du type **(29)** de la solution. On appellera *erreur absolue* à l'étape $(n + 1)$-ième la quantité : $e_{n+1} = x_{n+1} - x_{n+1} \qquad n = 0, \ldots, N - 1$.

Il est clair que l'erreur absolue dépend avant tout de l'erreur de troncature à l'étape $(n + 1)$. Que la méthode soit compatible ne signifie par pour autant que :

$$\textbf{(30)} \qquad \lim_{h \to 0} e = 0 \quad \text{avec} \quad e = \underset{0 < i < N}{\text{Max}} |e_i|.$$

En effet, l'erreur d'approximation commise à l'étape n se propage à l'étape $n + 1$; aussi l'on peut avoir simultanément :

$$\lim_{h \to 0} \tau = U \qquad \lim_{h \to 0} l = +\infty.$$

Dans un tel cas, la méthode sera dite divergente. Elle sera convergente si la condition **(30)** est réalisée.

Méthodes multisteps

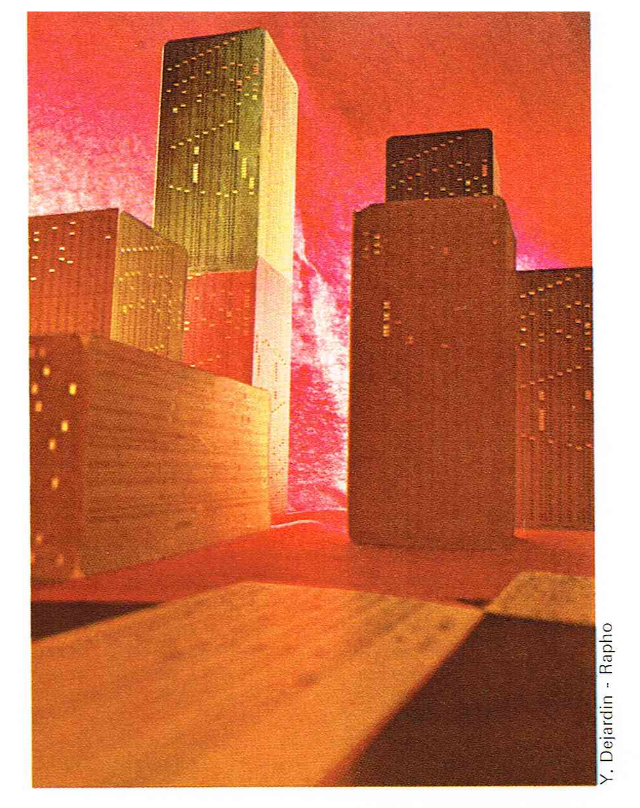
Les méthodes multisteps font appel à des fonctions F linéaires, c'est-à-dire de la forme :

$$\textbf{(31)} \qquad \tilde{x}_{n+1} = \sum_{i=0}^{r} a_i \tilde{x}_{n-i} + h \sum_{i=1}^{r} b_i f_{n-i}$$

où a_i et b_i sont des coefficients réels et

$$f_{n-i} = f(\tilde{x}_{n-i}, t_{n-i}).$$

On démontre que les méthodes multisteps sont *compatibles* et *convergentes*.



Y. Dejardin - Rapho

La **méthode « one step »** dite **méthode d'Euler** est la plus connue et la plus simple :

$$\tilde{x}_{n+1} = \tilde{x}_n + hf(\tilde{x}_n, t_n).$$

Écrite sous la forme :

$$\frac{\Delta \tilde{x}_{n+1}}{h} = f(\tilde{x}_n, t_n)$$

cette égalité n'est autre que la « discrétisation » de l'équation différentielle **(28)** (c'est-à-dire que l'on remplace la dérivée $\dfrac{dx}{dt}$ par la dérivée discrète $\dfrac{\Delta y_{n+1}}{n}$). Dans la définition des méthodes multisteps plus générales, on fait intervenir les valeurs prises lors des n étapes précédentes. On obtient alors les coefficients a_i et b_i de la formule **(31)** en écrivant la formule de Newton-Cotes. On peut en effet écrire l'équation différentielle **(28)** sous forme intégrale :

$$\textbf{(32)} \qquad x(t) = x(t_0) + \int_{t_0}^{t} f[x(\tau), \tau]\, d\tau$$

La **méthode de « différence centrale »** :

$$\tilde{x}_{n+1} = \tilde{x}_{n-1} + 2^h f_n$$

La **méthode d'Adams** :

$$\tilde{x}_{n+1} = \tilde{x}_n + \frac{h}{24}[55 f_n - 59 f_{n-1} + 37 f_{n-2} - 9 f_{n-3}]$$

Une **méthode du quatrième ordre** :

$$\tilde{x}_{n+1} = \tilde{x}_n + \frac{h}{2-4}(9 f_{n+1} + 19 f_n - 5 f_{n-1} + f_{n-2})$$

Notons que cette dernière méthode est une méthode implicite puisque f_{n+0} dépend de $\tilde{x}_{n+1}$. Il existe encore d'autres méthodes multisteps.

La **méthode d'Adams-Brashforth** est une combinaison de la méthode d'Adams et de la méthode du quatrième ordre que l'on vient de donner. Il existe encore d'autres méthodes multisteps.

En dehors des méthodes multisteps, mentionnons la **méthode** dite **de Runge-Kutta** dont le principe consiste à approximer la valeur prise par la dérivée en $\tilde{x}_{n+1}$ par un développement de Taylor de la fonction f en $\tilde{x}_n$. L'ordre de ce développement peut être aussi élevé que l'on souhaite ; sous certaines conditions, cette méthode est une méthode convergente.

BIBLIOGRAPHIE

ABADIE J., *Non Linear Programming*, North-Holland Publishing Company, Amsterdam, 1967. - DEMIDOVITCH B. et MARON I., *Éléments du calcul numérique*, Éditions Mir, Moscou, 1973. - FIACCO et Mc CORMICK, *Non Linear Programming*, John Viley and Sons, New York, 1968. - SEGALEN et JOUVENT, *la Programmation non linéaire*, Dunod, Paris, 1971.

LOGIQUE

Le terme « logique » est un des plus multivoques que la tradition philosophique nous ait légués. Il n'est guère plus précis que « métaphysique » ou « morale ». C'est pourquoi l'on est souvent tenté de le préciser soit en l'attribuant à quelque auteur de système, comme lorsqu'on parle de la logique d'Aristote, de la logique de Kant, ou de la logique de Hegel, voire de la logique de Husserl, soit en indiquant de quel type de logique il s'agit, comme lorsqu'on dit « logique modale », « logique intuitionniste », ou « logique plurivalente », etc.

Notre terme s'appliquait donc à des matières aussi diverses, parfois même franchement hétérogènes. Mais aujourd'hui on l'applique à une discipline qui est un genre de mathématique, et à travers elle, à un objet qui n'est précisément autre que la mathématique elle-même. Autrement dit, c'est une manière de parler mathématiquement des mathématiques, mais une manière qui a des exigences propres, c'est-à-dire qu'on ne l'obtient pas par simple extension du langage mathématique à un domaine nouveau.

Mais si le sujet concerné par la « logique mathématique » n'est pas des plus faciles à définir, les logiciens mathématiciens s'y arrêtent aussi peu que les mathématiciens eux-mêmes s'arrêtent sur l'objet de la mathématique. L'accord entier est réalisé sur des techniques que les débats philosophiques, toujours ouverts, n'entament point. Car, aussi intéressantes et profondes qu'elles puissent sembler, les techniques n'exigent en réalité que des langages qui soient compatibles avec elles, ce qu'on peut satisfaire sans se lier par un langage unique. En revanche, un minimum de points sont acquis dès le départ : comme, par exemple, que la logique mathématique étudie des langues formalisées, c'est-à-dire des objets reproduisant une structure mathématique, qu'elle ne prétend par conséquent nullement refléter toute la richesse de la langue naturelle, qu'elle n'investit pas non plus les modalités du « nécessaire », du « possible » ou du « réel » sur lesquels on n'est pas encore parvenu à une clarté suffisante malgré d'énormes progrès réalisés dans les dernières discussions à ce sujet, que la logique mathématique classique admet comme un principe logique sain le principe du tiers exclu, qu'elle ne refuse pas, moyennant les précautions requises, l'usage de l'infini dans les mathématiques classiques.

Giraudon

▶ *Le système à la fois le plus ancien et le plus représentatif du sens que nous donnons aujourd'hui au terme* « logique » *est la* théorie du syllogisme *créée par le philosophe* Aristote *(384-322 av. J.-C.).*

C'est dire que la logique mathématique s'est élaborée en restreignant le domaine de ses investigations et sa curiosité aux questions qu'elle pouvait résoudre mathématiquement ; c'est par des renoncements qu'elle est devenue une discipline rigoureuse, précise, et féconde. Comment ? C'est ce que nous allons exposer.

Histoire de la logique

Chacun sait que le système à la fois le plus ancien et le plus représentatif du sens que nous donnons aujourd'hui au terme « logique » est la *théorie du syllogisme* créée par le philosophe grec Aristote qui a vécu de 384 à 322 avant J.-C. On a tenté de remonter au-delà d'Aristote pour retrouver les éléments qui auraient mis ce dernier sur la voie de sa découverte. Mais toutes les recherches effectuées dans cette direction sont demeurées indécises. Le mieux qu'elles établissent est ce que nous savons déjà : que l'art de la discussion, le souci d'en formuler les principes, d'en articuler les degrés, de préciser toujours la conclusion à laquelle on veut arriver et le point duquel on part avaient déjà conduit à isoler certains principes du raisonnement logique et à former une habitude, consciemment cultivée, du sens logique. La première trace s'en trouve déjà dans le poème de Parménide, qui nie au nom de la cohérence logique tout ce qu'Héraclite prétendait enseigner à partir de l'expérience directe, opposant ainsi ce qui est matière d'opinion et ce qui doit relever de la science. C'est le commencement d'une dialectique philosophique fondée sur une méthode rigoureuse aboutissant, en mathématiques, et en mathématiques seulement, à des conclusions définitives, grâce au même procédé que Platon désigne par l'expression « par hypothèse » (ἐξ ὑποθέσεως), et qui vise à établir un rapport purement logique entre certaines propositions.

Cependant rien ne permet de dégager de l'œuvre de Platon un système de logique aussi explicite que la théorie aristotélicienne du syllogisme. Celle-ci ne forme qu'une partie de l'œuvre logique d'Aristote qui occupe plusieurs volumes groupés sous le nom d'*« Organon »* et comprenant :

— les *Catégories* qui traitent la théorie des termes ;

— *De l'interprétation*, consacré à la théorie des propositions ;

— les *Premiers Analytiques* ou théorie du syllogisme proprement dit ;

— les *Seconds Analytiques* ou théorie de la démonstration, c'est-à-dire du raisonnement scientifique ;

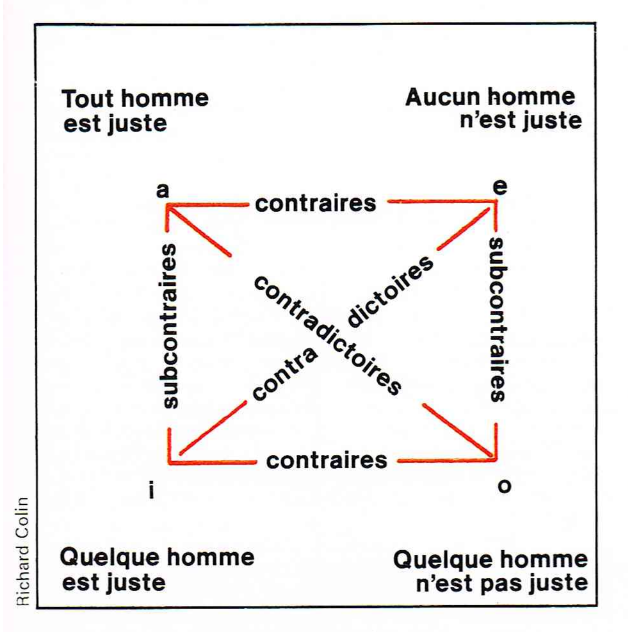
— les *Topiques* ou théorie du raisonnement dialectique ;

— *la Rhétorique* ou théorie du discours à effet oratoire.

L'ordre de ces traités ne reproduit pas celui de leur genèse historique, et les titres n'en sont pas, pour la plupart, le fait d'Aristote lui-même. Par exemple, Aristote a écrit les *Catégories* (de κατηγορεῖν qui signifie : attribuer quelque chose à quelqu'un ou quelque chose) et une partie des *Topiques* avant d'avoir élaboré sa théorie du syllogisme. C'est qu'il se proposait une enquête générale sur les conditions du discours et sur la nature de l'attribution ; mais, en affirmant que toute proposition revient à un jugement qui se compose d'un sujet et d'un attribut reliés par l'intermédiaire du verbe être, il doit préciser les diverses fonctions que peuvent remplir les termes : s'ils peuvent être, d'une part, sujets ou attributs, ils désignent, d'autre part, une substance (homme, cheval), ou bien indiquent le temps, le lieu, la qualité, la relation, etc.

Les *Premiers Analytiques* donnent d'abord une définition du syllogisme ; c'est, suivant Aristote, un discours « dans lequel certaines choses étant posées, une autre en résulte nécessairement par le seul fait que celles-là sont posées ». Apparemment, cette définition implique l'idée même d'inférence correcte, sous-jacente au concept d'hypothèse selon Platon, qui visait à préciser la notion de conséquence logique. Mais en fait, elle est plus restrictive : la définition des prémisses que donnent les *Premiers Analytiques* conduit à une classification élémentaire en quatre types ; « une prémisse est une proposition qui énonce quelque chose sur quelque chose d'autre de manière affirmative ou négative » (24 A 16). Sachant que toute proposition nous dit si un attribut appartient (B appartient à tout A) ou n'appartient pas (B n'appartient à aucun A) au tout, propositions universelles affir-

matives ou négatives, on peut avoir une partie seulement du sujet (B appartient à quelque A ou n'appartient pas à quelque A) ; d'où la division des propositions qui constituent les prémisses aussi bien que la conclusion d'un syllogisme en propositions universelles, affirmatives ou négatives, et particulières, affirmatives ou négatives, symbolisées depuis les *Scholastiques* par les lettres *a, e, i, o* (*a, i* pour les propositions affirmatives [de *affirmo*] ; *e, o* pour les négatives [de *nego*]. *SaP* équivaut à « tous les S sont P » ; *SiP* à « quelques S sont P » ; *SeP* à « nul S n'est P » ; *SoP* à « non tous les S sont P ») et ayant entre elles les rapports logiques qu'illustre le fameux carré :



Richard Colin

L'aspect peut-être le plus important de la syllogistique d'Aristote est l'idée axiomatique qui l'anime.

Aristote part des *syllogismes* dits *de la première figure,* qui sont les plus parfaits, c'est-à-dire les plus « évidents », pour démontrer les modes d'autres figures par application de certaines règles. Par exemple, le syllogisme en Barbara de la première figure : si A est affirmé de tout B, et B de tout C, A est nécessairement affirmé de tout C est un syllogisme parfait dont la conclusion est immédiate étant donné la hiérarchie entre les termes A, B, C.

D'autre part, le *syllogisme de la deuxième figure :* si M est affirmé de tout N mais d'aucun O, O n'est affirmé d'aucun N (syllogisme en Camestres qui se réduit à Celarent) ; mais « M n'est affirmé d'aucun O » équivaut à « O n'est affirmé d'aucun M », ce qui permet de retrouver le deuxième mode de la première figure. C'est là une méthode directe de réduction d'un syllogisme concluant à un syllogisme de la première figure : mais on ne peut l'appliquer en toute circonstance.

Considérons, par exemple, le syllogisme suivant (syllogisme en Baroco réduit par voie indirecte à Barbara) : si M est affirmé de tout N, mais non de quelque O, alors N peut ne pas être affirmé de quelque O, car, dit Aristote, s'il est affirmé de tout O et si M est affirmé de tout N, alors M est affirmé de tout O, contrairement à l'hypothèse qui nous dit qu'il n'est pas affirmé de quelque O. Ce qui suppose le recours à un principe logique non explicitement posé :

$$((P \wedge Q) \rightarrow R) \leftrightarrow ((P \wedge \neg R) \rightarrow \neg Q)$$

Aristote montre donc que tous les modes de toute figure se réduisent, directement ou indirectement, à ceux de la première figure. Il rassemble en tout 14 modes concluants et exclut, par la méthode du contre-exemple, toutes les autres possibilités.

La syllogistique d'Aristote montre donc à l'œuvre un esprit logique et axiomatique parfaitement conscient de ses tâches ; et c'est surtout le premier système logique qui utilise correctement des variables (de termes). Mais le champ d'application de cette logique est limité par la nature des propositions considérées ; de plus, certains principes logiques effectivement utilisés restent tout à fait implicites ; enfin, comme Aristote, dans les *Topiques* et les *Analytiques,* ne considère les propositions et les syllogismes que comme les « images » produites dans l'esprit des autres par certaines expressions verbales, il ne distingue point entre les mots et les choses, entre un nom et ce qu'il désigne.

Si Théophraste qui succède à Aristote (décédé en 322) à la tête du Lycée, n'apporte que des modifications de détail au système dont il a hérité, le premier stoïcisme, dont la formation doit beaucoup à Euclide de Mégare, dit le Socratique, et à son école, manifeste pour la logique un intérêt que le stoïcisme romain des Ier et IIe siècles perdra à peu près complètement, au profit de la morale. L'ancien stoïcisme, dont le centre est situé à Athènes, au IIIe siècle avant J.-C., s'est donc illustré, d'une part par les grands noms de Zénon de Cittium (336-264), de Cléanthes (331-232), de Chrysippe (280-210) à propos duquel on disait : « si les Dieux font de la dialectique, ils ne se servent pas d'une autre que de celle de Chrysippe », d'autre part par Philon, élève de Diodore, qui a défini le connecteur « si... alors » au sens extensionnel de l'implication matérielle.

On peut dire que les stoïciens, dont nous ne connaissons la doctrine que par ce que nous en ont rapporté Cicéron, Sextus Empiricus et Diogène Laërce, ont accompli une extension de la logique d'Aristote. Celui-ci utilisait « naïvement », pour prouver la validité ou l'invalidité des syllogismes, des lois logiques dont la syllogistique ne pouvait rendre compte ; pour expliciter cette base implicite il fallait une logique différente, à la fois par sa structure et par son intention. En effet, les stoïciens ne s'intéressent pas seulement à la question de savoir si un prédicat convient à un sujet, mais à des propositions susceptibles en général d'être vraies ou fausses. Il en résulte d'abord une distinction entre le signe, le sens et la dénotation : car une proposition peut être vraie ou fausse à condition de n'être pas confondue avec l'énoncé de la proposition qui n'est pas susceptible, lui, d'être vrai ou faux. Les stoïciens possédaient donc les rudiments d'une sémantique.

Ensuite, en cherchant la façon dont les propositions s'impliquent mutuellement, ils isolent les différents connecteurs propositionnels. Sextus Empiricus nous rapporte une sorte de table de vérité pour l'implication ; il nous dit qu'il y a quatre manières de combiner les éléments d'une implication : antécédent vrai et conséquent vrai, antécédent faux et conséquent vrai, antécédent vrai et conséquent faux, antécédent faux et conséquent faux ; l'implication n'étant fausse que si l'antécédent est vrai et le conséquent faux.

Enfin, les stoïciens utilisaient des *variables de proposition,* et ont développé une conception déductive de la logique des propositions. Ce qu'on peut bien voir dans les schémas d'inférence correspondant aux anapodictiques de Chrysippe, c'est-à-dire à des propositions qui n'ont pas besoin de démonstration :

(1) *Si le premier, le second*
Le premier
Donc le second

justifiant, par exemple, le célèbre raisonnement : « s'il fait jour, il fait clair ; or il fait jour, donc il fait clair ».

(2) *Si le premier, le second*
Non le second
Donc non le premier

illustré, par exemple, par « s'il fait jour, il fait clair ; or il ne fait pas clair, donc il ne fait pas jour ».

(3) *Non à la fois le premier et le second*
Or le premier
Donc non le second

par exemple : « il n'est pas vrai que Platon soit mort et vivant ; or Platon est mort, donc Platon n'est pas vivant ».

(4) *Le premier ou le second*
Or le premier
Donc non le second,

par exemple : « ou il fait jour, ou il fait nuit ; or il fait jour, donc il ne fait pas nuit ».

(5) *Le premier, ou le second*
Non le premier
Donc le second

par exemple : « ou il fait jour ou il fait nuit, or il ne fait pas jour, donc il fait nuit ».

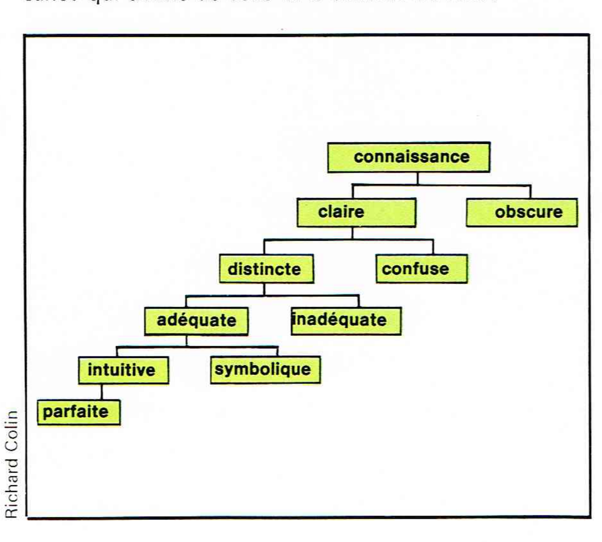
Mais ce ne sont là que les rudiments d'un système dont on est loin de connaître tous les éléments ; suffisamment riche pourtant pour avoir éveillé un sens logique parfaitement visible dans les efforts consacrés aux antinomies logiques dont celle du menteur : l'homme qui dit « je mens », s'il dit vrai, dit faux ; et s'il dit faux, dit vrai en même temps.

Parmi les diverses spéculations médiévales on ne trouve aucune innovation qui égale le système d'Aristote ou celui des stoïciens. En revanche la philosophie logique se développe avec Abélard (1079-1142), qui inaugure le problème des universaux, marque le point de départ de la méthode de discussion scolastique vulgarisée par les fameuses disputes, distingue entre l'implication vraie en vertu de sa forme et celle qui ne peut être vérifiée sans recours aux faits, analyse la copule « est » en vue de réduire tout énoncé catégorique à un énoncé de la forme « A est B ».

Ockham, Buridan, Albert de Saxe et le Pseudo Scot ont continué cette réflexion au cours du XIVe siècle, en l'orientant essentiellement vers l'élaboration d'une logique de la déduction qui a abouti à un certain nombre de lois logiques correspondant à des théorèmes classiques du calcul propositionnel ; par exemple, « toute partie d'une conjonction suit la conjonction dont elle fait partie », c'est-à-dire « $A \wedge B \rightarrow A$ » et « $A \wedge B \rightarrow B$ ».

On doit convenir que la logique « moderne » a très peu contribué à la formation de ce qu'on appelle aujourd'hui la logique. *La Logique ou l'Art de penser* d'Antoine Arnauld (1612-1694) et Pierre Nicole (1625-1695) a vulgarisé la distinction entre extension et compréhension d'un concept (entre l'ensemble des objets qui tombent sous lui et l'ensemble des attributs qui le composent). Mais on peut considérer que Joachim Jung et Arnold Genlinex, le premier par son influence directe sur Leibniz, le second par l'idée qu'il a maintenue d'une logique pure, ont perpétué l'exigence qui alimentera les différents projets leibniziens.

Gottfried Wilhelm Leibniz (1646-1716) est le premier qui soit allé plus loin dans le sens d'une logique comme langue artificielle susceptible de permettre la reconstruction de toute la pensée. C'est son projet d'une *caractéristique universelle* qui donne un contenu original à l'*Ars magna* de Raymond Lulle (1233-1315), constitue un début de réalisation de l'idée cartésienne de langue universelle, et inaugure une algébrisation de la logique dont le seul défaut est d'être restée ignorée. Leibniz, en effet, n'a pas publié le plus important de son œuvre logique qui ne vit le jour qu'après la naissance de la logique mathématique proprement dite, à la fin du XIXe siècle. Jusque-là, on citait Leibniz davantage pour sa *théorie de la connaissance* qui donne de celle-ci le schéma suivant :



que pour ses essais logiques, bien que l'idée de connaissance symbolique ne puisse être précisée que par une connaissance entièrement fondée sur une manipulation de symboles si réglée qu'elle semble aveugle, c'est-à-dire conforme à certaines règles fixées à l'avance et appliquées automatiquement, comme dans le *calcul de l'identité algébrique* $(a + b)(a - b) = a^2 - b^2$, où l'on pose :

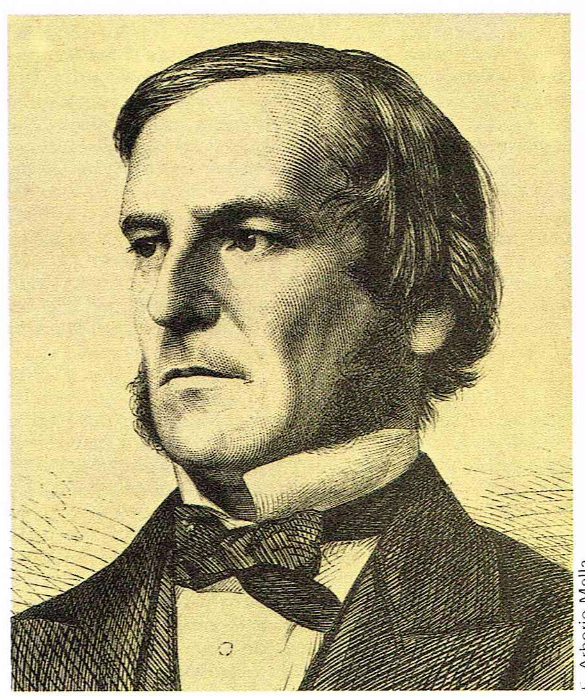
$$\begin{array}{r} a + b \\ a - b \\ \hline a^2 + ab \\ - ab - b^2 \\ \hline a^2 + 0 - b^2 \end{array}$$

Or l'idée d'établir la logique comme calcul résout l'antinomie inhérente à toute conception axiomatique de la logique puisque toute théorie axiomatique suppose déjà une logique ; les calculs au contraire n'en supposent pas.

Bernard Bolzano (1781-1848) s'adonne également à des recherches logiques d'autant plus intéressantes qu'elles sont l'œuvre d'un précurseur de la théorie des ensembles, soucieux de démonstrations mathématiques rigoureuses, purement « analytiques », ce qui signifie pour lui d'abord purement logiques. Dans son œuvre monumentale, la *Wissenschaftslehre,* ce qui retient l'attention, c'est surtout son élucidation du concept d' « analyticité » en étroite liaison avec l'idée de validité universelle. Une proposition est universellement valide suivant Bolzano si le résultat de toute substitution à un élément déterminé de cette proposition donne une proposition vraie ; elle est universellement non valide si toute substitution de ce genre donne une proposition fausse. L'analyticité d'une proposition par rapport à un de ses éléments constituants, c'est sa validité ou invalidité universelle par rapport à cet élément ; mais en un sens plus restreint, une proposition est analytique si elle est analytique eu égard à tous ses éléments constituants, à l'exclusion des « termes » logiques. Mais Bolzano ne dispose d'aucun critère pour distinguer ce qui est « logique » de ce qui ne l'est pas. En revanche, ses définitions de la non-contradiction d'un ensemble de propositions et de l'implication formelle sont toutes modernes et annoncent les méthodes d'Alfred Tarski.

La remise en chantier la plus considérable reste cependant celle de l'école anglaise, avec essentiellement George Boole (1815-1864) et Augustus de Morgan (1806-1878), créateurs de l'algèbre de la logique. Tous deux étaient attentifs aux similitudes de structure entre lois logiques et lois algébriques. Les efforts de Boole culminent dans l'axiomatisation de Huntington (1874-1952).

Le style algébrique de Boole, dont la seule différence avec Leibniz est, dit-on parfois, qu'il a publié ses travaux au contraire de celui-ci, ne prend les dimensions d'une véritable logique mathématique qu'avec l'œuvre de Richard Dedekind (1831-1916), dont le but explicite est de fonder les mathématiques sur une logique algébrique suffisamment développée pour répondre aux finesses de l'analyse. Dedekind est le seul auteur qui



▶ *A gauche, schéma de la* **théorie de la connaissance** *selon Leibniz. A droite,* **George Boole** *(1815-1864), l'un des créateurs de l'algèbre de la logique.*

donne, avant l'école italienne, toute son ampleur au style algébrique en logique, et cela grâce à sa théorie des systèmes qui est une application originale de l'algèbre de la logique au problème des fondements des mathématiques, dans toute la richesse qu'il a à la fin du XIXe siècle. On peut certes dire que c'est avec Dedekind, et Golttlob Frege (1848-1925) mais d'un autre point de vue, que naît la logique mathématique moderne, comme discipline propre, *non plus orientée exclusivement sur l'algébrisation de la logique qui préoccupait Boole, mais sur l'ensemble des problèmes rencontrés en mathématique.*

L'entreprise de Frege illustre remarquablement l'extension originale qu'est en train de prendre à ce moment notre discipline. On répétera, bien sûr, que Frege a été le premier à avoir élaboré un système axiomatisé du calcul des propositions, le premier à introduire les quantificateurs et surtout à en faire un emploi cohérent, le premier à concevoir une logique du second ordre avec quantification des variables de prédicat (ou de classe), mais son originalité profonde réside dans la façon dont il a fait l'analyse de certaines propositions qu'il a pris soin de « choisir » suffisamment complexes pour ne pas devoir chercher une extension, avérée impossible, des résultats obtenus sur des propositions simples à des propositions moins simples. Ayant profité du travail d'axiomatisation de l'arithmétique élémentaire, dû surtout à Hermann Grassmann (1809-1877), et des résultats d'Ernst Schröder d'une part, de Karl Weierstrass (1815-1897) de l'autre, il a pu isoler l'élément essentiel dont la logicisation emportait d'emblée celle de l'ensemble de l'arithmétique, et, par voie de conséquence, celle de l'analyse. Il s'agit de l'*induction mathématique*, que l'empirisme régnant réduisait soit à un raisonnement par analogie, soit, directement ou indirectement, à une induction empirique. Hermann Lotze (1817-1881), dont Frege suivait les cours à l'université de Göttingen, développa une critique vigoureuse de ces vues empiristes. Mais le mérite de Frege reste entier en ce sens qu'il a constitué une logique du second ordre indispensable à l'insertion du principe d'induction dans un corps de propositions purement logiques. Cela a naturellement conduit Frege à élaborer les principes du calcul des propositions, à faire une étude du concept d'égalité, à montrer l'usage de la quantification, à déployer, enfin, en étroite connexion avec les exigences de cette dernière, une conception sémantique de la logique, c'est-à-dire une logique liée à une théorie de la vérité, la plus élaborée qu'on connaisse avant la sémantique ensembliste de Tarski. Tout cela fait évidemment de Frege, non seulement le plus grand logicien de tous les temps, le Newton de la logique mathématique dont la fécondité est suffisamment attestée aujourd'hui par le génie de Gödel et de Tarski, mais aussi le Descartes de la philosophie contemporaine, sans lequel on ne comprendrait ni Wittgenstein, ni Carnap, de même qu'il serait impossible de comprendre Malebranche ou Leibniz sans référence à Descartes.

L'œuvre monumentale de Bertrand Russell (1872-1970) et Alfred Whitehead (1861-1947), les *Principia mathematica*, a tiré le plus grand bénéfice des travaux de Frege, malgré l'absence d'influence directe. Elle combine le style frégéen à l'héritage reçu de l'école de Boole et enrichi par la grande synthèse de Schröder. L'intérêt de ce monument vient de ce qu'il fut publié après la découverte du paradoxe dit de Russell (et/ou de Cantor) sur les ensembles qui se comprennent eux-mêmes comme éléments; aussi est-il devenu la référence des travaux axés sur la question des fondements des mathématiques et marqués par une double orientation : axiomatique, celle-ci ayant permis, grâce à Hilbert, de montrer l'unité profonde de la géométrie et de l'arithmétique ensembliste, la théorie des ensembles ayant elle-même, avec Zermelo, fait sienne la méthode hilbertienne. C'est dans cette perspective, et en réponse au défi intuitionniste, que s'est formulé le fameux programme de Hilbert, par rapport auquel doivent être compris les premiers résultats de Gödel sur la complétude de la logique du premier ordre, aussi bien que sur l'incomplétude de la logique du second ordre, et par suite de l'arithmétique élémentaire.

La logique mathématique

Actuellement en logique on ne peut se passer de langages formels. L'idée d'une formalisation du langage remonte à Leibniz, mais n'a été réalisée qu'avec la *Begriffsschrift* (1879) de Frege. Non seulement la langue naturelle n'est pas assez exacte pour convenir à l'expression des données mathématiques, mais encore elle peut donner lieu à de dangereuses méprises. Mais l'idée de langage formel implique celle de son interprétation, ou de ses modèles. La *théorie des modèles*, qui n'a été constituée qu'après la découverte de la logique mathématique, permet un exposé élémentaire très clair de la logique. La notion clef en est naturellement celle de *vérité*, dont la définition fait le lien entre le langage formel et son interprétation, par l'intermédiaire des modèles; en effet, définir la vérité d'une proposition suppose donnés, non seulement la proposition elle-même, mais aussi le modèle dans lequel on interprète celle-ci (voir la définition plus loin).



▲ **Alfred North Whitehead,** ***auteur, avec Bertrand Russell, des* Principia mathematica*; on leur doit également la construction de la logique symbolique.***

Une proposition A est vraie dans un modèle $\mathcal{M}$, si dans le modèle $\mathcal{M}$ on associe la valeur « vrai » à la proposition A; $\mathcal{M}$ constitue dans ce cas un modèle de la proposition A. Sinon A est fausse dans $\mathcal{M}$ et $\mathcal{M}$ n'est pas un modèle de A. Plus généralement $\mathcal{M}$ est un modèle d'un ensemble de propositions Σ si $\mathcal{M}$ est un modèle pour toute proposition de l'ensemble Σ.

Le calcul des propositions

Cette définition de la vérité étant préalablement posée, nous pouvons présenter succinctement le langage formel le plus simple, dénommé *calcul propositionnel*. Celui-ci est l'étude d'un ensemble d'énoncés simples tels que « Jean est sorti » ou « 2 + 2 = 4 » appelés *propositions atomiques* et d'énoncés plus complexes, obtenus en composant les énoncés simples à l'aide de mots de liaison appelés les **connecteurs propositionnels** comme « et », « ou », « si... alors », etc. Au niveau le plus intuitif cet ensemble d'énoncés peut être interprété comme un « monde possible » où toute proposition est *soit vraie soit fausse*, et où la valeur de vérité de toute proposition complexe ne dépend que de la valeur de vérité des propositions simples : on exprime cette dernière propriété en disant que les **connecteurs** considérés sont **extensionnels.** « Parce que » fournit un exemple simple d'une conjonction qui ne peut donner lieu à un connecteur propositionnel. En effet, étant donné quatre propositions, toutes supposées vraies : « X est mort », « X a mangé un mets empoisonné », « Il neige », « C'est l'été », par combinaison des deux premières on obtient une proposition vraie : « X est mort parce qu'il a mangé un mets empoisonné », tandis que la combinaison des deux dernières donne une proposition fausse : « Il neige parce que c'est l'été ».

On dénombre habituellement cinq connecteurs extensionnels : un *connecteur unaire* se rapportant à une proposition unique, la *négation*, notée $\neg$, et quatre *connecteurs binaires* reliant entre elles deux propositions, la *conjonction*, la *disjonction*, le *connecteur « si... alors »* et le *« si et seulement si »*, notés respectivement $\wedge$, $\vee$, $\Rightarrow$, $\Leftrightarrow$. Ce ne sont pas les seuls possibles mais les plus courants.

On définit chacun de ces connecteurs en indiquant la valeur de vérité prise par le composé le plus simple auquel il donne lieu, en fonction des diverses combinaisons possibles de valeurs de vérité (vrai et faux) attribuées aux propositions atomiques composantes. Ainsi la négation d'une proposition est vraie exactement quand cette proposition est fausse ; la conjonction de deux propositions n'est vraie que dans le seul cas où les deux propositions sont vraies, elle est fausse si l'une des deux ou les deux sont fausses ; la disjonction de deux propositions n'est fausse que dans le seul cas où les deux propositions sont fausses, elle est vraie si l'une des deux (ou les deux) est vraie ; l'implication dite matérielle, c'est-à-dire le connecteur désigné ci-dessus par « si... alors », n'est fausse que si la proposition antécédente est vraie et la conséquente est fausse ; elle est vraie si l'antécédent et le conséquent sont vrais, si l'antécédent et le conséquent sont faux, et si l'antécédent est faux et le conséquent vrai ; « si et seulement si » est vrai si les deux propositions sont simultanément vraies ou simultanément fausses, il est faux si l'une des deux est vraie tandis que l'autre est fausse.

On notera en passant que le $\vee$ du calcul propositionnel correspond à la *disjonction inclusive* de la langue usuelle, et que l'implication matérielle est définie en désaccord total avec l'usage intuitif usuel de l'expression « implique » qui désigne généralement une sorte de *connexion nécessaire* entre deux phénomènes. Une proposition comme « si $2 + 2 = 5$, alors la Seine traverse Paris » est une proposition vraie du point de vue du calcul propositionnel, tandis qu'elle n'a pas de sens dans la langue usuelle, vu qu'il n'y a aucun rapport entre l'antécédent et le conséquent.

De manière plus *formelle*, le langage du calcul des propositions P est constitué d'abord d'un certain nombre de **symboles primitifs**, c'est-à-dire non définis, qui sont :

— les *lettres de proposition : p, q, r*, etc., en nombre infini, qui représentent les propositions atomiques ;

— les *symboles de connecteurs :* $\neg$, $\wedge$, $\vee$, $\rightarrow$, $\leftrightarrow$.

Cette liste est redondante, car pour définir P, il suffit de $\neg$ et de l'un seulement des quatre connecteurs binaires cités les trois autres pouvant s'exprimer en fonction de celui-ci et de la négation. Nous choisirons ici le couple de connecteurs ($\neg$, $\wedge$). Dans certains cas, un seul connecteur suffit pour définir P ; cela arrive lorsque ce connecteur inclut en lui-même l'idée d'une négation, par exemple la barre de Scheffer, notée p/q, qui est équivalente à $\neg\, p \vee \neg\, q$;

— les *symboles de parenthèses :* (,).

Toute suite finie de symboles pris parmi les précédents est un **mot** ; parmi les mots on distingue ceux qu'on appelle les expressions bien formées ou **formules du calcul des propositions**, définies inductivement par les règles suivantes :

— toute lettre de proposition isolée est une formule ;

— si A est une formule, ($\neg$ A) est une formule ;

— si A et B sont des formules, (A $\wedge$ B) est une formule ;

— un mot est une formule seulement s'il est obtenu par un nombre fini d'applications des trois règles précédentes.

On peut également donner une définition récursive de la notion de **formule**, fondée sur la longueur des suites finies de symboles du vocabulaire initial : un symbole isolé est une formule si et seulement si c'est une lettre de proposition.

Une suite finie A de symboles de longueur $n > 1$ est une formule si et seulement s'il existe des formules B et C de longueur inférieure à n telles que A est soit ($\neg$ B) soit (B $\wedge$ C).

En termes ensemblistes, on peut encore dire (d'une troisième façon) que l'ensemble des formules F est le plus petit sous-ensemble de P contenant les lettres de proposition et tel que s'il contient les formules A et B, il contienne aussi ($\neg$ A) et (A $\wedge$ B).

Étant donné cette définition inductive de l'ensemble des formules, l'appartenance d'une propriété à toutes les formules ne pourra être établie que par induction ; pour une propriété φ on établira :

— que toute lettre de proposition a la propriété φ ;

— que si la formule A est de la forme ($\neg$ B) et que B ait la propriété φ, alors A a la propriété φ ;

— que si A est de la forme (B $\wedge$ C) et que B et C aient la propriété φ, alors A a la propriété φ.

Sémantique formelle de P

Rappelons que nous nous plaçons dans le cadre d'une logique bivalente, c'est-à-dire précisément où il y a deux valeurs de vérité : le *vrai*, représenté par le symbole 1, et le *faux* représenté par 0. Une lettre de proposition peut avoir une de ces deux valeurs.

Soit A une formule contenant n lettres de proposition, disons : $p_1, p_2, ..., p_n$. Assigner une valeur de vérité déterminée à chacune des lettres de proposition p_i, c'est définir une fonction δ définie sur l'ensemble $\{p_1, p_2, ..., p_n\}$ et à valeurs dans $\{0, 1\}$. La valeur de vérité de A est relative à une telle assignation, appelée souvent encore distribution de valeurs de vérité sur les lettres de proposition apparaissant dans A. Plus généralement on parlera de distribution de valeurs sur les lettres de proposition du langage P et on définira récursivement la notion de valeur de A pour une distribution donnée δ :

— si A se réduit à une lettre de proposition, sa valeur de vérité est déjà connue ;

— si A est de la forme $\neg$ B, la valeur de vérité de A est l'*opposée* de celle de B (0 si B a la valeur 1, 1 si B a la valeur 0) ;

— si A est de la forme (B $\wedge$ C), sa valeur est 1 si B et C ont toutes deux la valeur 1 ; sinon c'est 0.

Une formule A est dite *tautologie* si elle a la valeur 1 pour toute distribution de valeurs de vérité sur les lettres de proposition qu'elle contient : ce qu'on note $\vdash$ A.

La valeur d'une formule A pour une distribution donnée se calcule facilement de proche en proche : partant de la valeur assignée aux lettres de proposition, on calcule progressivement la valeur des sous-formules de A, de la plus simple à la plus complexe, pour arriver enfin à A. On peut réunir sous forme de tableau les résultats indiquant la valeur de vérité de A pour toutes les distributions de valeurs possibles sur les lettres de proposition qu'elle contient ; pour n lettres on a 2^n possibilités (il faut se souvenir ici qu'on a défini une distribution de valeurs de vérité sur n lettres comme une fonction :

$$\{0, 1, 2, ..., n\} \rightarrow \{0, 1\},$$

et connaître par ailleurs le résultat élémentaire de la théorie des ensembles selon lequel l'ensemble de telles fonctions est de cardinal 2^n). Ce tableau s'appelle la **table de vérité** de A.

Par exemple, prenons pour A la formule :

$$p \wedge (\neg\, q \wedge \neg\, (p \wedge r)).$$

p q r	¬q	p∧r	¬(p∧r)	¬q∧¬(p∧r)	A
1 1 1	0	1	0	0	0
1 1 0	0	0	1	0	0
1 0 1	1	1	0	0	0
1 0 0	1	0	1	1	1
0 1 1	0	0	1	0	0
0 1 0	0	0	1	0	0
0 0 1	1	0	1	1	0
0 0 0	1	0	1	1	0

Richard Colin

▶ *La* table de vérité *de A.*

Ce procédé simple (devenu courant en logique essentiellement depuis les travaux de Post en 1921), quoique parfois assez long, fournit un moyen mécanique et fini pour déterminer si une formule est une tautologie : on écrit sa table de vérité et on vérifie si on trouve la valeur 1 à toutes les lignes de la table. Notons en passant qu'une

formule qui n'est pas une tautologie peut être soit une *formule neutre*, vraie pour certaines distributions de valeurs sur ses lettres de proposition, fausse pour d'autres, ainsi que c'est le cas de notre formule A précédente, soit une *antilogie* (ou *contradiction*) si elle est fausse pour toute distribution de valeurs. Notons surtout que ce procédé fournit une définition formelle très commode des connecteurs propositionnels que nous avons présentés de façon intuitive plus haut :

p	¬p	p	q	p ∧ q	p ∨ q	p → q	p ↔ q
1	0	1	1	1	1	1	1
0	1	1	0	0	1	0	0
		0	1	0	1	1	0
		0	0	0	0	1	1

Richard Colin

Nous venons de dégager la notion cardinale de tautologie ; soulignons qu'il s'agit là d'une *notion syntaxique*, en ce sens qu'elle ne fait pas appel aux modèles. D'un point de vue *sémantique* on peut définir formellement (par induction encore une fois) l'expression « A est vraie dans un modèle $\mathcal{M}$ », en entendant par modèle simplement un sous-ensemble de P ; on notera $\mathcal{M} \models$ A. Une *formule* A sera alors *valide* si elle est vraie dans tous les modèles de P.

Le **premier théorème de complétude** établit l'équivalence logique des concepts de validité et de tautologie et autorise par conséquent l'utilisation des tables de vérité pour déterminer si une formule est ou non valide. Ce premier théorème sera suivi d'un autre du même nom mais de plus large portée. Symboliquement $\vdash$ A si et seulement si $\models$ A ; on remarquera que le signe $\vdash$ s'emploie dans un contexte syntaxique et le signe $\models$ dans un contexte sémantique. Ce premier théorème établit en fait leur équivalence.

Définition du concept de déduction formelle

Il faut introduire d'abord dans notre langage P la **règle d'inférence** connue sous le nom de *modus ponens* qui établit que de « A » et « A → B » on peut conclure à « B ». Puis on dira que A est *déductible* d'un ensemble de formules Σ, symboliquement : $\Sigma \vdash$ A, si et seulement s'il existe une suite finie $B_0, B_1, ..., B_n$ de formules telle que A soit précisément B_n et que toute formule B_i soit, ou bien un élément de l'ensemble Σ, ou bien une tautologie, ou bien obtenue par application du *modus ponens* à deux formules antérieures dans la suite. La suite $B_0, B_1, ..., B_n$ constitue une déduction de A à partir de Σ. Notons que A est déductible de l'ensemble vide si et seulement si c'est une tautologie. La propriété la plus importante de la déduction est énoncée par le **théorème de la déduction :**
si $B_0, B_1, B_2, ..., B_n \vdash$ A *alors* $B_0, B_1, B_2, ..., B_{n-1} \vdash B_n \rightarrow$ A.

Le théorème de complétude élargi

Un ensemble Σ *de formules est dit inconsistant si et seulement si* $\Sigma \vdash$ *A pour toute formule A.*

Par ailleurs nous avons déjà vu que $\mathcal{M}$ est un modèle de Σ si et seulement si toute formule de Σ est vraie dans $\mathcal{M}$. Σ est dit satisfaisable si et seulement s'il a au moins un modèle. Le **théorème** suivant, le plus important du calcul des propositions, donne un critère pour déterminer quand un ensemble de formules est satisfaisable : Σ *est consistant si et seulement s'il est satisfaisable.* Comme le précédent ce théorème jette un pont entre le point de vue syntaxique et le point de vue sémantique.

Un corollaire de nature purement sémantique est connu sous le nom de **théorème de compacité :** Σ *est satisfaisable si tout sous-ensemble fini de* Σ *est satisfaisable.*

Le théorème de compacité a de nombreuses applications dont nous ne pouvons parler ici.

Le calcul des prédicats du premier ordre

Une phrase comme « tous les nombres sont plus grands que 1 » constitue dans le calcul propositionnel une proposition atomique qu'on ne cherche pas à analyser plus précisément. Nous allons maintenant présenter un langage plus riche où cette analyse sera possible ; le calcul des prédicats étudie, en effet, les expressions du type : « tous », « il y a », « quelques », etc., et ce qu'on appelle *prédicats*, ou mieux et pour les désigner par leur nom mathématique usuel aujourd'hui, les *relations* mettant en jeu un ou plusieurs individus, comme : « est rouge », « plus grand que », « appartient à », ou même « est » qui s'emploie en des sens bien différents. Les noms des individus entrant dans la relation considérée sont désignés comme des *constantes d'individu ;* le nombre des individus auxquels s'applique la relation s'appelle le *nombre de places* de la relation ou du prédicat.

Par exemple « est un homme » dans « Socrate est un homme » est un prédicat à une place ; « plus grand que » est un prédicat à deux places ; « *sept* est la somme de *trois* et *quatre* » nous donne l'exemple d'un prédicat à trois places. Plus généralement on parlera de relations à *n* places où *n* est un nombre entier positif.

Formellement les prédicats seront notés par des majuscules latines P, Q, ... et P (x) signifiera : « x a la propriété P ». L'expression « pour tout... » est appelée *quantificateur universel*, notée $\forall$; « il y a », « il existe », « quelques » sont synonymes et s'appellent *quantificateur existentiel*, noté $\exists$. « Tout x a la propriété P » sera donc noté : $\forall x$ P (x) ; « il existe un x ayant la propriété P » sera noté : $\exists x$ P (x). On sait que $\exists x$ P (x) est logiquement équivalent à $\neg \; \forall x \neg$ P (x), en sorte qu'on peut se limiter, pour définir formellement L, au seul symbole du quantificateur universel.

Dans les expressions « $\forall x$ P (x) » et « $\exists x$ P (x) », x s'appelle une *variable liée* ou *muette ;* ce qui signifie que le sens de l'expression n'est pas changé si on substitue partout à x, y ou toute autre minuscule latine. « $\forall x$ P (x) » a la même signification que « $\forall y$ P (y) ».

Définition formelle du langage L

Le langage L est déterminé par deux groupes de symboles : les *symboles de relation* et les *symboles de constantes d'individus.* Pour compléter la liste des symboles primitifs, il faut ajouter :

— les symboles : x_0, x_1, x_2, ..., appelés *variables d'individus ;*
— les *connecteurs propositionnels*, disons $\neg$ et $\wedge$;
— le *quantificateur universel* $\forall$;
— les *parenthèses gauche* et *droite* (,).

Toute suite finie de symboles pris dans cette liste est un mot de L, et on définit l'ensemble des formules ou expressions bien formées par induction, d'une manière analogue à celle déjà employée pour le calcul propositionnel :

(1) les formules atomiques sont de la forme P $(x_1, x_2, ..., x_n)$ où P est un prédicat à *n* places et les x_i des variables d'individus ;

(2) si A et B sont des formules, ($\neg$ A) et (A $\wedge$ B) sont des formules ;

(3) si x est une variable d'individu et A une formule, $\forall x$ A est une formule ;

(4) une suite de symboles n'est une formule que si elle est obtenue par un nombre fini d'applications des règles **(1)**, **(2)** et **(3)**.

Variables libres et variables liées

Supposons que $\forall x$B est une sous-formule d'une formule A ; B est appelée la *portée* du quantificateur universel $\forall x$. Toute occurrence de x dans $\forall x$B est une *occurrence liée* de x dans A. Une occurrence non liée de x dans A est une *occurrence libre.*

Soit, par exemple, dans la formule :

$$P\ (x, y) \rightarrow \forall x\ (\forall y\ R\ (x, y) \rightarrow Q\ (x, y))$$

La première occurrence de x est libre, les trois autres occurrences de x sont liées, les première et dernière occurrences de y sont libres, la seconde et la troisième sont liées ; la formule $\forall y$ R $(x, y) \rightarrow$ Q (x, y) constitue la portée du quantificateur universel $\forall x$, et R (x, y) celle du quantificateur universel $\forall y$.

Une variable est libre dans une formule A si elle y a au moins une occurrence libre dans A. Une *formule* est *close* si aucune des variables qu'elle contient n'y est libre. Une formule close est une proposition en ce sens qu'elle est soit vraie, soit fausse, tandis qu'à une *formule non close* ou *ouverte*, appelée *forme propositionnelle*, on

ne peut associer de valeur de vérité déterminée. Par exemple, dans l'arithmétique usuelle, la forme propositionnelle « $x < 5$ » est vraie pour certains entiers, fausse pour d'autres ; la proposition « $\forall x\ (x < 5)$ » est fausse et la proposition « $\exists x\ (x < 5)$ » est vraie. On aura remarqué au passage qu'il suffit de quantifier la variable libre x dans « $x < 5$ » pour transformer la forme propositionnelle en proposition, « $x < 5$ » étant vraie pour certains entiers, on dira qu'elle est *réalisable* ou *satisfaisable* dans $\mathbb{N}$. Mais elle n'est pas *valide* dans $\mathbb{N}$ puisqu'elle n'est pas vraie pour tous les éléments de $\mathbb{N}$.

Si x est une variable d'individu et z une variable ou une constante d'individu, si A est une formule quelconque, S^x_z A désigne la formule obtenue en substituant z à toutes les occurrences libres de x dans A.

Modèles

Soit X un ensemble et k un entier positif. Un k-tuple d'éléments de X est une suite ordonnée d'éléments de X ; l'ensemble de tous les k-tuples de X est noté X^k. Un prédicat P à k-places sur X est un ensemble de k-tuples formés d'éléments de X, c'est-à-dire un sous-ensemble de X^k ; on dit que les éléments de ce sous-ensemble *satisfont* P. Par exemple, si X est l'ensemble des entiers naturels, l'ensemble : $\{(x, y) \in X^2 ; x < y\}$ est une relation binaire (à 2 places) satisfaite par le couple (3,4) mais non par le couple (4,3).

Un *monde possible* ou modèle $\mathcal{M}$ du langage L se constitue d'une part d'un ensemble X, ou *univers*, non vide ; d'autre part d'une fonction $\mathfrak{J}$ qui interprète toute lettre de prédicat à k-places comme une relation à k-places sur X, et toute constante d'individu comme un élément de X ; $\mathcal{M} = (X, \mathfrak{J})$. On remarquera que la situation est beaucoup plus compliquée qu'en calcul des propositions, car ici les propositions ou formules closes énoncent quelque chose sur les *individus* du modèle considéré ; il n'est plus aussi aisé de décider si une proposition de langage L est vraie ou fausse dans tel modèle, alors qu'on disposait d'un moyen de résoudre la même question pour toute proposition du langage P et pour tout modèle de P.

Plus précisément considérons une proposition A de L et cherchons si A est vraie dans le modèle $\mathcal{M}$; si A est de la forme $\neg$ B ou encore de la forme B $\wedge$ C, il est clair que nous saurons répondre à la question si nous savons si B et C sont vraies dans $\mathcal{M}$; mais si A est de la forme $\forall x$ B, il se peut que x soit libre dans B et par conséquent B n'est plus une formule close alors que A l'était. Dans ce dernier cas cela n'a pas de sens de demander si B est vraie dans $\mathcal{M}$; tout ce que nous pouvons demander, c'est si pour chaque élément a en particulier de l'ensemble X constituant le domaine de base de $\mathcal{M}$, B est vraie ou non. Mais supposons que B soit elle-même de la forme $\forall y$ C, on retrouve la même difficulté : on ne pourra poser la question de la vérité de C que pour un couple d'éléments a, b de X. Plus généralement, si on convient de noter A $(x_0, x_1, ..., x_n)$ une formule A de L où les variables libres de A forment un sous-ensemble de $\{x_0, x_1, ..., x_n\}$, on demandera si A est vraie lorsqu'on prend le n-uplet $(a_0, a_1, ..., a_n)$ d'éléments de X pour les variables

$$x_0, x_1, ..., x_n.$$

Nous ne pouvons développer ici les moyens qui permettent de résoudre proprement cette difficulté ; ceux-ci conduiraient évidemment à une définition par induction de la relation $\mathcal{M} \models$ A, qui signifie que la proposition A est vraie dans $\mathcal{M}$ ou que $\mathcal{M}$ satisfait A ou encore que $\mathcal{M}$ est un modèle de A. Mais pour permettre une comparaison avec ce qui se passe dans le calcul propositionnel, nous pouvons présenter cette définition par récurrence dans une situation simplifiée : on considérera qu'une formule atomique est de la forme P $(a_1, a_2, ..., a_n)$ où P est une lettre de prédicat et les a_i des constantes, éléments de X. Dans ce cas :

1) $\mathcal{M} \models$ P $(a_1, a_2, ..., a_n)$ si $(a_1, a_2, ..., a_n)$ satisfait l'interprétation (P) de P dans $\mathcal{M}$;

2) $\mathcal{M} \models \neg$ B si et seulement si non $\mathcal{M} \models$ B ;

3) $\mathcal{M} \models$ B $\wedge$ C si et seulement si $\mathcal{M} \models$ B et $\mathcal{M} \models$ C ;

4) $\mathcal{M} \models \forall x$ A si et seulement si pour tout $a \in X$, $\mathcal{M} \models S^x_a$ A.

Par exemple, si X est l'ensemble des nombres naturels et P la relation $<$, on a $\mathcal{M} \models \forall x\,(P\,(6, x) \rightarrow P\,(3, x))$, mais non $\mathcal{M} \models \forall x\,(P\,(3, x) \rightarrow P\,(6, x))$. On dira que A est universellement valide si $\mathcal{M} \models$ A pour tout modèle $\mathcal{M}$ de L.

Maintenant, dans le cas où A n'est pas une proposition ou formule close, c'est-à-dire, avec la convention adoptée plus haut si A est de la forme A $(x_1, x_2, ..., x_n)$, on dira qu'elle est universellement valide si et seulement si sa clôture universelle $\forall x_1 \forall x_2 ... \forall x_n$ A $(x_1, x_2, ..., x_n)$ est valide.

Étant donné un ensemble Σ de propositions, $\mathcal{M}$ est un modèle de Σ, si et seulement si $\mathcal{M}$ est un modèle pour toute proposition appartenant à Σ. Une proposition ou un ensemble de propositions est *satisfaisable* (ou *réalisable*) si et seulement s'il a au moins un modèle.

Une proposition A est *conséquence* d'une proposition B si et seulement si tout modèle de B est aussi un modèle de A.

Il reste à définir la notion de *déduction formelle* avant de citer les théorèmes importants du calcul des prédicats : le théorème de complétude et le théorème de compacité. La situation est tout à fait analogue à celle qu'offrait le calcul des propositions. Considéré comme système formel, L se compose d'un certain nombre d'axiomes et de règles d'inférence.

Les axiomes sont de deux types :

— les axiomes propositionnels, toute tautologie T de calcul des propositions peut donner lieu à un axiome de L quand on substitue convenablement aux lettres de propositions contenues dans T des formules de L ;

— les axiomes concernant l'usage des quantificateurs ; par exemple si A et B sont des formules de L et x une variable non libre dans A, alors la formule $\forall x\,(A \rightarrow B) \rightarrow (A \rightarrow \forall x\,B)$ est un axiome logique.

Quant aux règles d'inférence, on en admet deux :

— le *modus ponens* déjà utilisé en calcul des propositions ;

— la *règle de généralisation propre* ou *calcul des prédicats :* de A on peut inférer $\forall x$ A.

Moyennant ces données, la notion de déduction est absolument la même qu'en calcul des propositions : ajoutons simplement qu'on appelle **théorème** une proposition qui se déduit des axiomes donnés, à l'aide des règles d'inférence admises. « A est un théorème » s'écrit : si Σ est un ensemble de propositions, $\Sigma \vdash$ A signifie qu'il existe une déduction (ou preuve) de A à partir des axiomes et de Σ. On écrit en général simplement $\Sigma \vdash$ A, étant naturellement sous-entendu que les axiomes font également partie des hypothèses. Σ est dit *inconsistant* si toute formule de L se déduit de Σ et *consistant* dans le cas contraire. On peut énoncer alors :

— Le **théorème de complétude de Gödel :** *toute proposition A est un théorème de L si et seulement si A est (universellement) valide. C'est-à-dire $\vdash$ A si $\models$ A.* D'où l'équivalence logique de la notion syntaxique de déduction et de la notion sémantique de conséquence.

— Le **théorème de complétude élargi :** *soit Σ un ensemble de propositions ; Σ est consistant si et seulement s'il a un modèle.*

— Le **théorème de compacité :** *un ensemble de propositions Σ a un modèle si et seulement si tout sous-ensemble fini de Σ a un modèle.*

Conclusion

Ce sont là quelques-uns des concepts fondamentaux de la logique mathématique qui connaît encore actuellement un développement des plus intenses. Elle n'a pas seulement réalisé tous les programmes logiques anciens en leur donnant l'exactitude et la concision voulues, en écartant des disciplines rigoureuses ambiguïtés et malentendus issus d'une maîtrise insuffisante des moyens d'expression, en précisant l'idée de fondement des mathématiques non seulement par l'élaboration explicite des principes utilisés le plus souvent implicitement, mais encore des concepts de déduction et de démonstration ; bref, en réalisant cette idée surgie peu à peu dans les mathématiques modernes, qu'il ne suffit pas d'accumuler des résultats, qu'il est indispensable pour une science rigoureuse de savoir où elle va et comment elle y va. Il n'est donc pas étonnant que les résultats les plus impressionnants de la logique mathématique soient des résultats négatifs : ils nous disent ce qu'il est vain de chercher et nous orientent sur ce qu'il est toujours possible, pour ainsi dire, *a priori*, de trouver.

Univac

▲ *L'informatique prend une importance croissante par la généralisation de l'emploi des ordinateurs, instruments à la fois de travail et de recherche.*

INITIATION A L'INFORMATIQUE

Généralités

L'informatique, considérée comme science du traitement rationnel de l'information, support des connaissances et des communications, n'est pas une discipline très récente. Cependant, les progrès technologiques ayant permis la réalisation de machines de plus en plus complexes, l'informatique prend une importance croissante dans presque tous les secteurs d'activité, par la généralisation de l'emploi des ordinateurs, instruments à la fois de travail et de recherche.

La machine de Pascal constitue en quelque sorte le point de départ du développement des machines à calculer. Il s'agit d'une machine à roues dentées fonctionnant suivant le principe encore utilisé aujourd'hui des compteurs totalisateurs mécaniques. Chaque roue comporte dix positions, portant les chiffres de 0 à 9. Lors du passage de 9 à 0, la roue fait avancer d'une position la roue suivante. L'opération de base, sur une telle machine, est l'addition. On peut également réaliser des multiplications par additions successives portant sur les puissances de 10 successives.

Le développement de cette technique a abouti aux machines à calculer mécaniques effectuant les quatre opérations (et quelquefois des opérations plus complexes, comme l'extraction de racines carrées) et imprimant les résultats. Si les performances de telles machines représentent un grand pas en avant par rapport à la machine de Pascal, les principes de base et le mode d'emploi n'en sont pas moins les mêmes. Il faut introduire manuellement les données, et actionner les touches d'opérations dans l'ordre voulu pour obtenir le résultat. De plus, il est souvent nécessaire de noter des résultats intermédiaires, soit pour les réintroduire dans une phase ultérieure du calcul, soit pour prendre des décisions sur la façon de mener la suite du calcul.

Pour automatiser ces opérations il faut donc que le calculateur puisse stocker des résultats intermédiaires, et soit capable d'enchaîner les opérations et, éventuellement, de modifier l'enchaînement en fonction de certains résultats.

Par exemple, si on cherche les solutions réelles d'une équation du second degré,

$$ax^2 + bx + c = 0$$

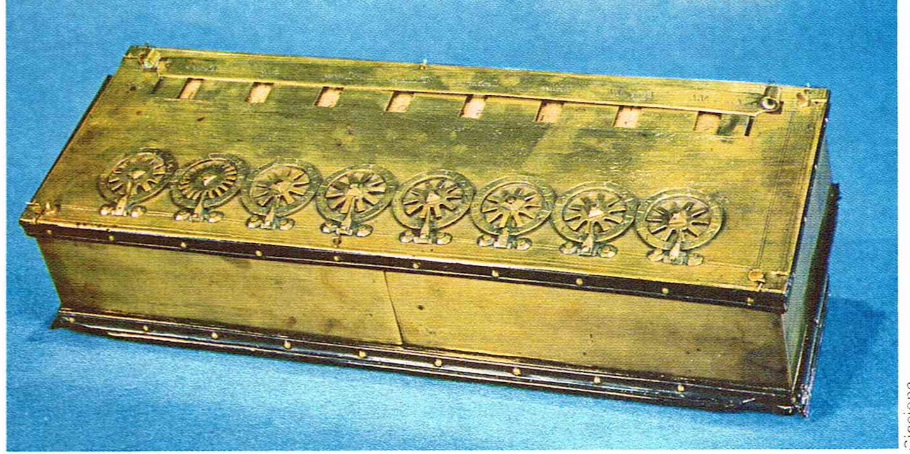
on calcule d'abord : $\Delta = b^2 - 4\,ac$;
ensuite, si $\Delta < 0$, il faut arrêter le calcul, car dans ce cas l'équation n'a pas de solution réelle;
si $\Delta \geqslant 0$, on peut calculer les solutions :

$$x_1 = \frac{-b - \sqrt{\Delta}}{2\,a}, \qquad x_2 = \frac{-b + \sqrt{\Delta}}{2\,a}$$

Bien entendu, il est cependant nécessaire d'introduire les données ainsi que la suite d'opérations à effectuer et les conditions dans lesquelles les effectuer.

C'est à peu près ce que réalisaient les machines mécanographiques, associées à un bloc de calcul. Ces machines étaient munies d'un bloc de commande câblé et pouvaient

▼ *La « machine arithmétique » de Pascal (Paris, musée des Arts et Métiers).*



Ciccione

effectuer des calculs et des opérations tels que des tris, en utilisant des cartes perforées comme support de l'information.

Par ailleurs, le mathématicien anglais Babbage avait, au milieu du XIXe siècle, conçu un type de machine capable de recevoir les données d'un problème et la suite d'opérations à effectuer pour arriver au résultat souhaité *(Analytical Engine)*. Certains principes se retrouvent dans les calculateurs actuels, mais la technologie entièrement mécanique de l'époque et la complexité de la machine n'en ont jamais permis la réalisation complète.

Deux éléments importants ont permis l'évolution vers les machines actuelles.

Le *premier est technologique :* le développement de l'électronique a permis la réalisation de machines de plus en plus performantes en capacité, en rapidité et en complexité.

Le *deuxième est théorique :* c'est l'introduction par *Goldstine* et *von Neumann,* en 1947, de la notion de programme enregistré.

L'idée consiste à stocker ensemble le programme et les données dans la mémoire de la machine, sans faire *a priori* de distinction sur leur nature. La distinction se fait par la façon logique de les structurer, non par la façon physique de les introduire dans la machine.

Avec l'apparition de la notion de programme enregistré, la structure générale des machines s'est à peu près fixée. La technologie et les méthodes d'utilisation des calculateurs, en particulier les moyens de communication avec la machine, continuent d'évoluer grâce aux progrès techniques d'une part, aux travaux théoriques d'autre part.

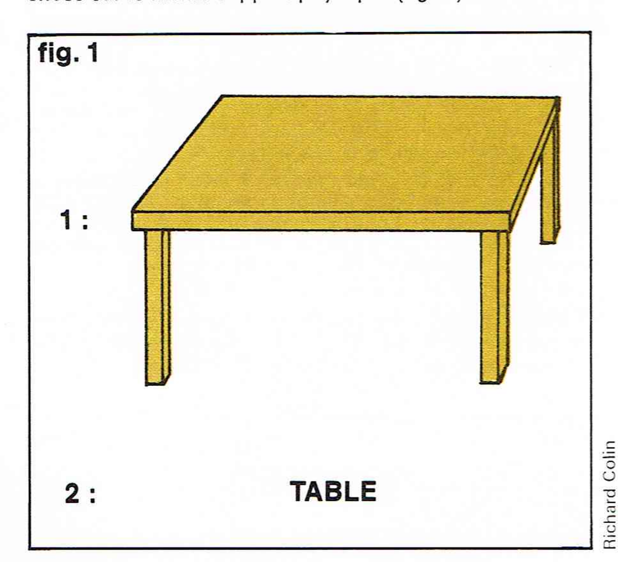
Représentation de l'information

Pour pouvoir traiter une information, il est nécessaire que cette information soit sous une forme accessible à l'organe de traitement, c'est-à-dire que cet organe puisse « comprendre » l'information. Cela suppose deux choses :

— un support physique de l'information ;

— un mode de représentation (« support logique ») sur le support physique.

Voici deux représentations différentes d'une même chose sur le même support physique (*fig.* 1) :



▶ ***Figure 1 :* un exemple de deux représentations différentes d'une même chose sur le même support physique.**

Dans les deux cas, le support physique est constitué de traits d'encre sur du papier. Dans le premier cas le mode de représentation est graphique, figuratif ; dans le deuxième cas, c'est une représentation à l'aide d'un mot de la langue française.

On peut de la même façon, prendre le mot TABLE et le faire supporter par des supports physiques différents, par exemple :

— comme précédemment, on peut l'écrire ;

— on peut le prononcer et l'enregistrer à l'aide d'un magnétophone. Dans ce cas, le support physique de l'information est la bande magnétique.

Dans ce qui suit, nous allons essentiellement traiter des modes de représentation de l'information plutôt que des supports physiques.

Représentation des nombres

Les nombres sont représentés à l'aide de symboles appelés chiffres. Il existe plusieurs modes de représentation, ou systèmes de numération.

● *Les systèmes non pondérés :* chaque symbole a une valeur intrinsèque, ne dépendant pas de l'emplacement du symbole dans la représentation. C'est le cas de la numération romaine.

● *Les systèmes de position :* une pondération est introduite par la position du symbole. Une catégorie importante est constituée par les systèmes à base constante. Dans un système à base b (b est un nombre entier), on représente les nombres à l'aide de b symboles : $S_0, S_1, ..., S_{b-1}$ représentant les valeurs entières de 0 à $b - 1$. Ainsi, dans le système décimal ($b = 10$) on utilise les chiffres 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. On représente les nombres avec ces symboles et éventuellement un point pour séparer la partie entière de la partie fractionnaire (l'usage en informatique est d'utiliser le point et non la virgule, conformément à l'écriture anglo-saxonne des nombres ; compte tenu de l'importance croissante de l'usage des ordinateurs, et pour éviter des notations multiples pouvant créer des confusions, l'utilisation du point tend à se généraliser dans les autres disciplines).

La représentation suivante :

$$a_n\, a_{n-1} \ldots a_1\, a_0 \cdot a'_1\, a'_2 \ldots a'_p \qquad \text{(I)}$$

désigne le nombre :

$$N = a_n\, b^n + a_{n-1}\, b^{n-1} + \ldots a_1\, b + a_0 + a'_1\, b^{-1} + a'_2\, b^{-2} + \ldots + a'_p\, b^{-p} \qquad \text{(II)}$$

Tous les a, a' étant des symboles pris parmi $S_0, S_1, ..., S_{b-1}$.

Pour simplifier les notations, nous avons confondu les symboles et les valeurs qu'ils représentent. (I) est une représentation symbolique, alors que (II) est une expression arithmétique. Nous verrons qu'il convient d'être prudent lorsqu'on fait ce genre de confusion (voir *Adresse symbolique*).

Passage dans une base donnée

Considérons le nombre précédent ; nous allons traiter séparément la partie entière, notée E(N) et la partie fractionnaire F(N).

Posons $N_1 = E(N)$, on a :

$$N_1 = a_n\, b^n + a_{n-1}\, b^{n-1} + \ldots + a_1\, b + a_0.$$

On peut mettre b en facteur dans les n premiers termes :

$$N_1 = b\,(a_n\, b^{n-1} + a_{n-1}\, b^{n-2} + \ldots + a_1) + a_0. \quad (\text{avec } a_0 < b)$$

Sous cette forme, on voit que a_0 est le reste de la division entière de N_1 par b, le quotient étant :

$$N_2 = a_n\, b^{n-1} + a_{n-1}\, b^{n-2} + \ldots + a_1.$$

De même, a_1 est le reste de la division entière de N_2 par b.

En recommençant l'opération jusqu'à l'obtention d'un quotient nul, on obtient successivement les valeurs correspondant aux chiffres de la représentation de N dans le système à base b.

Exemple

Trouver la représentation en système octal (base 8) du nombre décimal 253

253 : 8 = 31 reste 5
31 : 8 = 3 reste 7
3 : 8 = 0 reste 3

La représentation octale est donc 375.

Posons maintenant $M_1 = F(N)$, on a $M_1 < 1$:

$$M_1 = a'_1\, b^{-1} + a'_2\, b^{-2} + \ldots + a'_p\, b^{-p};$$

calculons $b \times M_1$:

$$b \times M_1 = a'_1 + a'_2\, b^{-1} + \ldots + a'_p\, b^{-p+1}$$

on voit que a'_1 est la partie entière de $b \times M_1$.

Posons $M_2 = F(b \times M_1) = a'_2\, b^{-1} + \ldots a'_p\, b^{-p+1}$; de même a'_2 est la partie entière de $b \times M_2$.

En continuant l'opération, on obtient les différents chiffres de la représentation de M_1 dans le système à base b. Il faut remarquer que cette représentation n'est pas toujours finie. Dans la démonstration précédente, nous avions supposé *a priori* une représentation finie. De plus une partie fractionnaire à représentation finie dans une base peut avoir une représentation infinie dans une autre.

Exemple

La fraction $F = \frac{1}{3}$ est représentée en décimal par $F = 0.3333...$ avec une infinité de 3 et par $F = 0.1$ en système à base 3.

Autre exemple

Trouver la représentation en système à base 5 du nombre décimal 0.5.

$0.5 \times 5 = 2.5$ 2 partie entière, .5 partie fractionnaire

$0.5 \times 5 = 2.5$, etc.

La représentation cherchée est donc 0.2222...

Nous venons de voir deux algorithmes (algorithme : suite d'opérations à effectuer) permettant de convertir, d'un système de numération à un autre, les parties entières et fractionnaires d'un nombre. Remarquons que dans les exemples donnés, il s'agissait de passer du système décimal à un autre système; les calculs étaient effectués en décimal, c'est-à-dire dans le système de départ. On peut également faire la conversion en effectuant les calculs dans le système d'arrivée. Il suffit alors, dans ce système, de calculer l'expression (**II**) où les a et a' désignent les chiffres de la représentation de départ et b la base de départ. Reprenons un exemple précédent à l'envers :

convertir N = 375 (OCTAL) en décimal,

$$N = 3 \times 8^2 + 7 \times 8 + 5 = 253 \text{ (DÉCIMAL)}$$

Les systèmes habituellement utilisés en informatique (voir tableau I) sont les suivants.

— *Système décimal :* utilisé essentiellement dans les opérations d'entrée de données ou de sortie de résultats car c'est le système auquel tout le monde est habitué.

— *Système binaire :* c'est dans ce système que travaillent les calculateurs électroniques.

— *Système octal.*

— *Système hexadécimal :* système à base 16. On utilise les lettres A, B, C, D, E, F, comme chiffres représentant les valeurs 10, 11, 12, 13, 14, 15.

L'intérêt des systèmes octal et hexadécimal est que la conversion du binaire dans l'un de ces systèmes et l'inverse sont immédiats. Cette propriété est d'ailleurs vraie quand deux systèmes sont tels que la base de l'un est une puissance entière de la base de l'autre.

En effet, reprenons l'expression :

$$N_1 = a_n b^n + a_{n-1} b^{n-1} + ... + a_p b^p + a_{p-1} b^{p-1} + ... + a_1 b + a_0$$

et divisons par b^p (division entière) :

$$N_1 : b^p = a_n b^{n-p} + a_{n-1} b^{n-1-p} + ... + a_p;$$

$$\text{reste} = a_{p-1} b^{p-1} + a_{p-2} b^{p-2} + ... + a_1 b + a_0$$

Les p derniers chiffres de N_1 : $a_p a_{p-1} ... a\ a_0$ représentent donc dans le système de base b la valeur du dernier chiffre dans le système de base b^p. Donc, pour passer du système à base b au système à base b^p, on groupe les chiffres par groupes de p en commençant par le chiffre des unités, et on remplace chaque groupe par le chiffre ayant même valeur dans le système de base b^p.

Le passage inverse se fait en remplaçant chaque chiffre par sa valeur exprimée sur p chiffres dans la base b.

Exemples

Passer de la représentation binaire 10110100 à la représentation octale : 10110100 → 264 (OCTAL).

Passer de la représentation hexadécimale A74 en binaire :

A → 1010, 7 → 0111, 4 → 0100 soit 101001110100.

L'octal et l'hexadécimal constituent des notations abrégées plus faciles à utiliser que le binaire, en en étant cependant très proches.

— *Système décimal codé binaire* (DCB) : il s'agit d'une représentation mixte. Le système de numération est décimal mais au lieu d'utiliser les chiffres décimaux, on les remplace par le nombre binaire exprimé sur quatre bits (**bit :** chiffre binaire, abréviation de l'anglais, **B***inary* D*igi***t**).

Exemple : 48 en décimal;

4 s'écrit 0100, 8 s'écrit 1000, 48 s'écrit 01001000.

— *Codes binaires pondérés :* en général, ils sont destinés à coder les chiffres décimaux. Ils s'expriment au moyen de 4 bits, chacun ayant un poids correspondant à sa position.

Tableau I - Représentation des premiers nombres entiers dans différents systèmes de numérotation.

Décimal	Binaire	Octal	Hexadécimal	DCB
1	1	1	1	0001
2	10	2	2	0010
3	11	3	3	0011
4	100	4	4	0100
5	101	5	5	0101
6	110	6	6	0110
7	111	7	7	0111
8	1000	10	8	1000
9	1001	11	9	1001
10	1010	12	A	0001 0000
11	1011	13	B	0001 0001
12	1100	14	C	0001 0010
13	1101	15	D	0001 0011
14	1110	16	E	0001 0100
15	1111	17	F	0001 0101
16	10000	20	10	0001 0110
17	10001	21	11	0001 0111
18	10010	22	12	0001 1000
19	10011	23	13	0001 1001
20	10100	24	14	0010 0000
21	10101	25	15	0010 0001
22	10110	26	16	0010 0010
23	10111	27	17	0010 0011
24	11000	30	18	0010 0100
25	11001	31	19	0010 0101
26	11010	32	1A	0010 0110

▲ *Tableau I :* **les systèmes de numération *habituellement utilisés en informatique.***

Exemple : le code 2421. Dans ce code 1110 représente 2 + 4 + 2 = 8. Il faut pouvoir coder tous les nombres de 0 à 9 avec les poids disponibles (ainsi, 6211 ne permet pas de coder 5). Remarquons que le code DCB est un code pondéré particulier : code 8421.

Les codes redondants

Les parasites électriques peuvent affecter les transmissions entre organes de traitement et faire qu'un bit soit mal détecté (0 au lieu de 1, ou inversement). Dans ce cas, il existe des méthodes permettant de détecter et éventuellement de corriger les erreurs. Elles consistent à ajouter des bits supplémentaires, qui ne sont pas significatifs du point de vue de l'information considérée (ces bits sont dits « redondants »), mais permettent de vérifier s'il y a eu erreur ou non.

— *Codes à parité fixe*

On ajoute un bit, appelé bit de parité, tel que le nombre total de bits égaux à 1 soit pair (respectivement impair), si on a choisi de travailler en parité paire (respectivement impaire). Si une erreur de transmission s'est produite, portant sur un bit, un contrôle de parité nous avertit de l'erreur, car la parité a été changée.

Exemple

En DCB, 8 et 9 s'écrivent respectivement 1000 et 1001. Ajoutons un bit de parité impaire en dernière position. On obtient respectivement 10000 et 10011. Si en cours d'opération, on trouve un code tel que 10001, on sait qu'il y a erreur, car ce code n'a pas de parité impaire. Bien entendu, un tel procédé ne permet que de détecter une erreur et non de la corriger. Il ne permet pas non plus de détecter deux erreurs simultanées, car, dans ce cas, la parité reste inchangée.

Tableau II - Code de Hamming impair obtenu à partir du code DCB

Décimal	DCB	Bit n° 1 (1, 3, 5, 7)	Bit n° 2 (2, 3, 6, 7)	Bit n° 4 (4, 5, 6, 7)	Code de Hamming résultant
0	0000	1	1	1	1101000
1	0001	0	0	0	0000001
2	0010	1	0	0	1000010
3	0011	0	1	1	0101011
4	0100	0	1	0	0100100
5	0101	1	0	1	1001101
6	0110	0	0	1	0001110
7	0111	1	1	0	1100111
8	1000	0	0	1	0011000
9	1001	1	1	0	1110001
Bits n°	3, 5, 6, 7				

▲ ***Tableau II :*** ***type de*** **code autocorrecteur.**

— *Codes autocorrecteurs (codes de Hamming)*

Dans le cas où une erreur peut provoquer des conséquences inacceptables, il existe un procédé de codage permettant de localiser une erreur, donc de la corriger puisque le caractère est binaire. Supposons qu'un code contienne initialement n bits. On va chercher, en ajoutant p bits de parité à ce code, à déterminer, en cas d'erreur, quel est le bit erroné. Le code obtenu comporte en tout $n + p$ bits ; l'information recherchée, en cas d'erreur, sera le numéro du bit erroné et doit être fournie par les p bits supplémentaires. On doit donc avoir $2^p \geqslant n + p + 1$.

En effet, on considère les cas suivants :

1 : Pas d'erreur.
2 : Erreur sur le premier bit.
3 : Erreur sur le second.
⋮
$n + p$: Erreur sur le $n + p$ — 1-ième bit.
$n + p + 1$: Erreur sur le $n + p$-ième bit.

Soit en tout, $n + p + 1$ cas différents.

On peut donc *a priori* dresser une table du nombre minimal de bits supplémentaires, en fonction du nombre de bits du code initial :

n	1	2	3	4	5	6	...	11	12	13	...	26
p	2	3	3	3	4	4		4	5	5		5

Nous allons mettre en place un tel code, en prenant par exemple : $n = 4$ donc $p = 3$. Ces trois bits seront des bits de parité, portant chacun sur une partie seulement du code. En vérifiant la parité de chaque groupe, nous pouvons fabriquer un nombre binaire de 3 bits, chaque bit valant 0 si la parité du groupe correspondant est correcte, 1 si elle est incorrecte. Cherchons à déterminer ces groupes, de telle façon que le nombre obtenu soit le numéro du bit erroné s'il y a erreur et 000 dans le cas où le code est correct.

Écrivons les valeurs binaires correspondant aux nombres de 1 à 7 :

	c_1	c_2	c_3
1	0	0	1
2	0	1	0
3	0	1	1
4	1	0	0
5	1	0	1
6	1	1	0
7	1	1	1

On voit qu'un contrôle de parité doit porter sur le groupe 4 5 6 7, pour déterminer le bit c_1, puisque celui-ci doit être à 1 si et seulement si une erreur s'est produite dans ce groupe ; un contrôle doit porter sur 2 3 6 7 pour déterminer c_2 et sur 1 3 5 7 pour déterminer c_3. Donc, il faut un bit de parité dans chacun de ces groupes. Les bits 1, 2 et 4 ne figurent chacun que dans un groupe. Ce sont donc ceux-là que l'on va choisir comme bits de parité, les bits 3 5 6 7 représentant le code initial. Pour éviter les confusions, on attribue aux bits les numéros correspondant à leur place dans le code résultant bien que le raisonnement précédent ne repose que sur la notion de numéro attribué au bit (ainsi, on pourrait numéroter les bits dans l'ordre 7 4 6 1 5 2 3 au lieu de 1 2 3 4 5 6 7).

Le *tableau* II montre un code de Hamming impair obtenu à partir du code DCB.

On peut vérifier l'effet de correction sur un exemple :
Soit le code de 5 : 1001101
changeons un bit : 1001001
Parité 1 3 5 7 : paire. Erreur, $c_1 = 1$
Parité 2 3 6 7 : impaire. Correct, $c_2 = 0$
Parité 4 5 6 7 : paire. Erreur, $c_3 = 1$
d'où $c_1c_2c_3 = 1\ 0\ 1$, soit 5 en décimal ; c'est bien le 5e bit qui avait été changé.

Ce procédé permet donc de corriger une erreur ; cependant, s'il y a deux erreurs, il est impossible de s'en rendre compte. Pour cela, on peut ajouter un bit de parité supplémentaire portant sur la totalité des bits. En cas d'erreur double, les contrôles de parité partiels ne seront pas satisfaits, alors que le contrôle total sera satisfait.

Ces divers procédés relèvent d'une notion qui est la notion de distance dans un code. Sans préciser exactement de quoi il s'agit, disons simplement, en raisonnant de manière intuitive, que plus les messages (ou mots) possibles d'un code sont différents entre eux, plus on a de chances de pouvoir corriger des erreurs, en choisissant d'interpréter systématiquement un mot erroné comme provenant du mot qui lui ressemble le plus dans le code : on suppose en effet que les erreurs multiples susceptibles de changer complètement l'aspect du message sont beaucoup moins probables que les erreurs simples n'apportant que de légères modifications.

Nous voyons qu'il est théoriquement possible de s'affranchir des erreurs, mais en pratique on est conduit à utiliser des codages plus longs que si on était sûr de ne pas avoir d'erreurs. Ceci a des conséquences économiques non négligeables. En effet, pour une capacité donnée de stockage, transmission ou traitement de l'information, on utilise une partie de cette capacité et du temps de traitement pour faire des contrôles sur de l'information redondante. Il est donc souhaitable de chercher à réaliser du matériel assez fiable pour n'avoir à faire que des contrôles simples (par exemple, contrôle de parité), en sachant que ces contrôles couvrent pratiquement tous les risques d'erreurs pouvant subsister. Cette condition augmente évidemment les coûts d'études du matériel, mais permet de réduire le coût d'exploitation.

Éléments de logique

Physiquement, les calculateurs électroniques ne réalisent que des opérations logiques. Nous verrons que certaines conventions permettent d'effectuer certaines opérations telles que des opérations arithmétiques, à l'aide d'opérations logiques.

Une variable logique est une variable ne pouvant prendre que deux valeurs : vrai ou faux. Conventionnellement, on représente ces valeurs respectivement par 1 ou 0, mais il ne faut pas attribuer à cette représentation un sens arithmétique (on pourrait aussi bien utiliser les symboles V et F ou + et —, etc.).

Une fonction logique de variables logiques est une fonction qui, pour toute combinaison des valeurs des variables, prend soit la valeur vrai, soit la valeur faux. Soit n variables logiques indépendantes, il existe 2^n combinaisons différentes des valeurs de ces variables, puisque chacune peut prendre deux valeurs ; pour chaque combinaison, une fonction logique peut prendre deux valeurs, il existe donc 2^{2^n} fonctions logiques différentes de n variables logiques.

On peut représenter une fonction logique à l'aide d'une table de vérité, en plaçant dans une colonne les combinaisons des variables, et dans une autre les valeurs correspondantes de la fonction.

Fonctions d'une variable

Il y en a quatre différentes (*tableau* III) :

Tableau III - Fonctions d'une variable.							
F_0		F_1		F_2		F_3	
Var.	F.	Var.	F.	Var.	F.	Var.	F.
0	0	0	0	0	1	0	1
1	0	1	1	1	0	1	1

La seule intéressante est la fonction F_2, appelée complémentation. On la représente par $F_2(a) = \bar{a}$, la barre indiquant la complémentation.

Fonctions de deux variables

Il y en a 2^{2^2}, soit 16. Certaines ont des propriétés intéressantes pour l'application pratique dans la réalisation des circuits logiques électroniques. Nous allons en examiner quelques-unes (*tableau* IV) :

Tableau IV - Table de vérité des fonctions de deux variables.						
Variables	F_1	F_6	F_7	F_8	F_{13}	F_{14}
0 0	0	0	0	1	1	1
0 1	0	1	1	0	1	1
1 0	0	1	1	0	0	1
1 1	1	0	1	0	1	0

F_1 : fonction intersection, ou fonction ET
Notation $F_1(a, b) = a \wedge b$ ou $a \cdot b$

F_6 : fonction OU EXCLUSIF
Notation $F_6(a, b) = a \oplus b$

F_7 : fonction réunion, ou fonction OU
Notation $F_7(a, b) = a \vee b$ ou $a + b$

F_8 : fonction exclusion ou fonction NI
Notation $F_8(a, b) = a \downarrow b$
Cette fonction est complémentaire de F_7, on peut donc écrire $F_8 = \overline{F_7}$. Pour cette raison on adopte plus souvent la notation $F_8 = \overline{a \vee b}$ ou $\overline{a + b}$

F_{13} : fonction implication
Notation $F_{13}(a, b) = a \Rightarrow b$

F_{14} : fonction incompatibilité
Notation $F_{14}(a, b) = a \mid b$
On adopte plus souvent la notation $F_{14} = \overline{a \cdot b}$ ou $\overline{a \wedge b}$
On appelle aussi cette fonction NON-ET
En effet, on voit que $F_{14} = \overline{F_1}$

Les fonctions 1, 6, 7, 8, 14 sont commutatives.
Les fonctions 1, 6, 7 sont associatives, c'est-à-dire :

$(a \cdot b) \cdot c = a \cdot (b \cdot c)$ on peut donc écrire $a \cdot b \cdot c$
$(a \vee b) \vee c = a \vee (b \vee c) = a \vee b \vee c$
$(a \oplus b) \oplus c = a \oplus (b \oplus c) = a \oplus b \oplus c$

Les fonctions ET (1) et OU (7) sont distributives l'une par rapport à l'autre, ce qui s'exprime par :

$$a \cdot (b \vee c) = (a \cdot b) \vee (a \cdot c)$$
$$a \vee (b \cdot c) = (a \vee b) \cdot (a \vee c)$$

Elles ont également les propriétés suivantes, qu'il est facile de vérifier en s'aidant des tables de vérité :

— Propriété d'idempotence : $a \cdot a = a$
$a \vee a = a$
— Propriété d'absorption : $a \cdot (a \vee b) = a$
$a \vee (a \cdot b) = a$
— Élément neutre : $a \cdot 1 = a$
$a \vee 0 = a$

1 et 0 sont respectivement éléments neutres des fonctions ET et OU.

La fonction OU EXCLUSIF (F_6) possède la propriété suivante :

$$a \oplus 1 = \bar{a}$$
$$a \oplus 0 = a$$

— Théorème de Morgan :

$$\overline{a \cdot b} = \bar{a} \vee \bar{b}$$
$$\overline{a \vee b} = \bar{a} \cdot \bar{b}$$

◄ *Tableau III : les 4 fonctions d'une variable (Var = variable, F = fonction).*

Décomposition canonique des fonctions logiques

Soit F $(a, b, c, ...)$ une fonction de plusieurs variables, on peut l'écrire sous la forme suivante :

$$F(a, b, c...) = a \cdot F(1, b, c...) \vee \bar{a} \cdot F(0, b, c...)$$

car si a vaut 1, $\bar{a}$ vaut 0 et seul F $(1, b, c...)$ compte, de même si a vaut 0 seul F $(0, b, c...)$ compte.

F $(1, b, c...)$ et F $(0, b, c...)$ peuvent se décomposer de la même façon. Par exemple, prenons le cas d'une fonction de trois variables F (a, b, c) :

$$F(a, b, c) = a \cdot F(1, b, c) \vee \bar{a} \cdot F(0, b, c)$$

On peut continuer :

$$F(1, b, c) = b \cdot F(1, 1, c) \vee \bar{b} \cdot F(1, 0, c)$$
et
$$F(0, b, c) = b \cdot F(0, 1, c) \vee \bar{b} \cdot F(0, 0, c)$$

En décomposant de même les quatre nouveaux termes et en reportant dans F (a, b, c), on obtient :

$$\begin{aligned} F(a, b, c) = {} & a \cdot b \cdot c \cdot F(1, 1, 1) \vee a \cdot b \cdot \bar{c} \cdot F(1, 1, 0) \\ & \vee a \cdot \bar{b} \cdot c \cdot F(1, 0, 1) \vee \bar{a} \cdot b \cdot c \cdot F(0, 1, 1) \\ & \vee a \cdot \bar{b} \cdot \bar{c} \cdot F(1, 0, 0) \vee \bar{a} \cdot b \cdot \bar{c} \cdot F(0, 1, 0) \\ & \vee \bar{a} \cdot \bar{b} \cdot c \cdot F(0, 0, 1) \vee \bar{a} \cdot \bar{b} \cdot \bar{c} \cdot F(0, 0, 0) \end{aligned}$$

En remplaçant F (0, 0, 0), F (0, 0, 1), etc., par les valeurs tirées de la table de vérité, la fonction se réduit à une réunion d'intersections dans lesquelles interviennent toutes les variables, soit sous forme directe, soit sous forme complémentée. Ces intersections sont appelées intersections de base.

Exemple

Soit la fonction de trois variables définie par sa table de vérité :

a	b	c	F
0	0	0	0
0	0	1	1
0	1	0	0
0	1	1	1
1	0	0	1
1	0	1	0
1	1	0	1
1	1	1	0

On peut écrire F de la façon suivante :

$$F = \bar{a} \cdot \bar{b} \cdot c \vee \bar{a} \cdot b \cdot c \vee a \cdot \bar{b} \cdot \bar{c} \vee a \cdot b \cdot \bar{c}$$

Cette façon d'exprimer une fonction logique à partir des intersections est appelée première forme normale ou première forme canonique. Il en existe une seconde qui consiste à exprimer la fonction par une intersection de réunions, appelées réunions de base.

Il nous suffit de la première pour remarquer que toutes les fonctions logiques, quel que soit le nombre de variables, peuvent s'exprimer en utilisant seulement les fonctions ET, OU et complémentation. En fait, on aurait pu introduire ces opérations de façon axiomatique. En effet, un ensemble muni d'opérations possédant les propriétés des fonctions ET, OU, complémentation, décrites ci-dessus, possède une structure d'algèbre de Boole.

◄ *Tableau IV : table de vérité des fonctions de deux variables.*

▶ ***Les machines à calculer électroniques de poche ne comportent souvent qu'un seul circuit intégré.***

D'un point de vue pratique, on dit que ces trois fonctions constituent un groupe complet, car elles permettent la réalisation de toutes les fonctions logiques. Les fonctions $\overline{a \cdot b}$ et $\overline{a \vee b}$ sont particulièrement intéressantes, car chacune forme à elle seule un groupe complet; on peut en effet réaliser les fonctions ET, OU et complémentation en utilisant l'une ou l'autre de ces deux fonctions comme le montrent les relations suivantes, faciles à vérifier :

$\bar{a} = \overline{a \cdot a}$	$a \cdot b = \overline{\overline{(a \cdot b)} \cdot \overline{(a \cdot b)}}$	$a \vee b = \overline{\overline{(a \cdot a)} \cdot \overline{(b \cdot b)}}$
$\bar{a} = \overline{a \vee a}$	$a \cdot b = \overline{\overline{(a \vee a)} \vee \overline{(b \vee b)}}$	$a \vee b = \overline{\overline{(a \vee b)} \vee \overline{(a \vee b)}}$

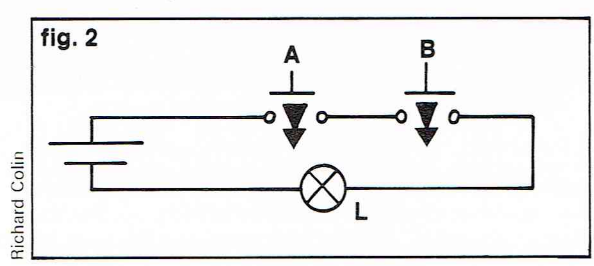
Réalisation des fonctions logiques

Les premières réalisations furent à base de contacts électriques et de relais. Deux contacts placés en série dans un circuit constituent un opérateur ET, car ils doivent être tous les deux fermés pour que le circuit électrique soit fermé. Deux contacts en parallèle constituent un opérateur OU. Un contact « à ouverture » ouvrant le circuit lorsqu'il est actionné constitue un opérateur de complémentation.

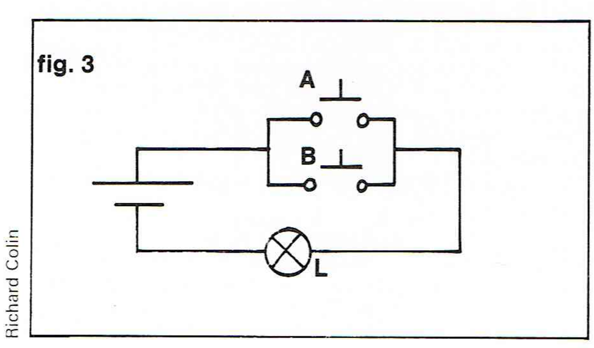
Exemples

Quand A *et* B sont actionnés, la lampe L s'allume. On peut écrire L = A · B (*fig.* 2).

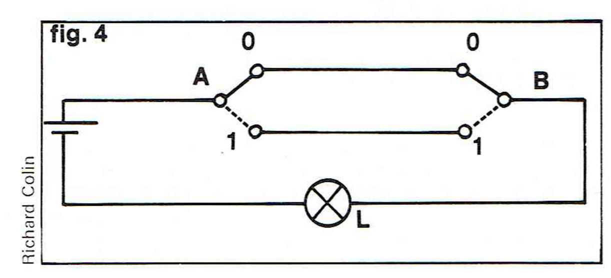
▶ ***Figures 2, 3, 4 : trois exemples de réalisation des fonctions logiques à l'aide d'éléments électromécaniques.***



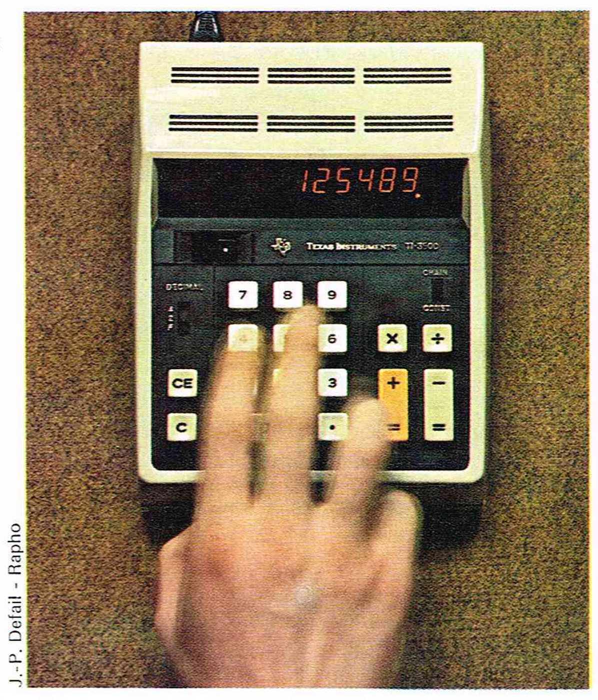
Quand A *ou* B (ou les deux) est actionné, la lampe L s'allume. On peut écrire L = A ∨ B (*fig.* 3).



Une commande en « va-et-vient » d'après la *figure* 4 peut être représentée par la fonction $L = A \cdot B \vee \bar{A} \cdot \bar{B}$.



Les premiers calculateurs électroniques furent réalisés avec des tubes électroniques (diodes, triodes, thyratrons, etc.). Le développement des transistors, puis des circuits intégrés a fait considérablement évoluer la technologie. En effet, les fabricants de circuits intégrés offrent actuellement des gammes très étendues de circuits, depuis

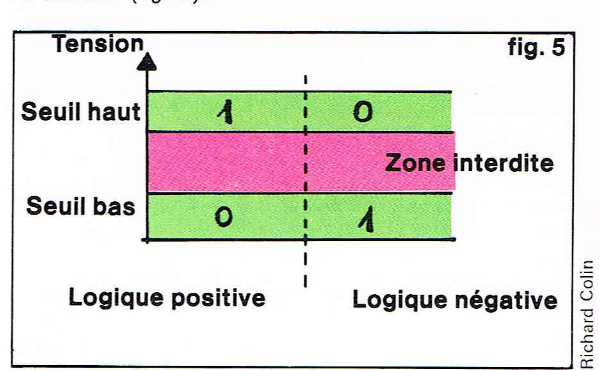


J.-P. Defail - Rapho

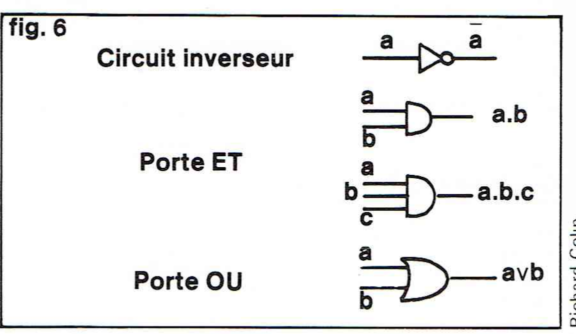
les plus simples jusqu'aux plus complexes (les machines à calculer électroniques de poche ne comportent souvent qu'un seul circuit intégré). Pour un ingénieur ayant à concevoir un organe d'ordinateur, ces circuits constituent des « boîtes noires », car il n'est pas nécessaire de connaître en détail comment ils sont faits, mais plutôt comment on doit les utiliser.

Ces circuits ont deux états électriques possibles. Un état correspond à une tension électrique supérieure à un seuil haut, l'autre à une tension inférieure à un seuil bas. Les circuits sont conçus de telle façon qu'ils ne peuvent pas prendre un état stable entre les seuils haut et bas en utilisation normale. Deux conventions sont possibles pour attribuer une valeur logique à ces niveaux électriques. En logique positive, le 1 logique est représenté par le niveau haut, en logique négative, le 1 est représenté par le niveau bas (*fig.* 5).

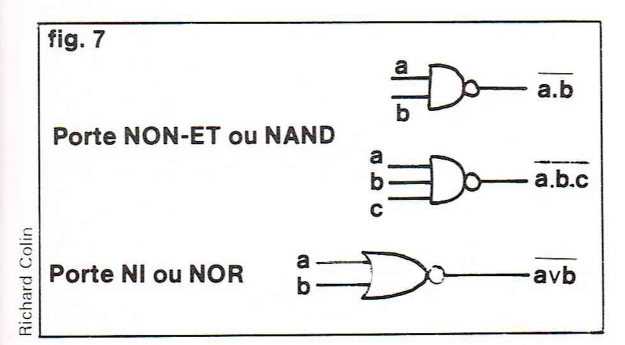
▶ ***Figure 5 : représentation schématique des états électriques des circuits logiques.***



Les noms donnés aux circuits correspondent aux fonctions réalisées en logique positive. Voici les circuits simples les plus courants et leurs représentations.



En logique négative, une porte ET réalise en fait une fonction OU et réciproquement (voir Théorème de Morgan) [*fig.* 6].



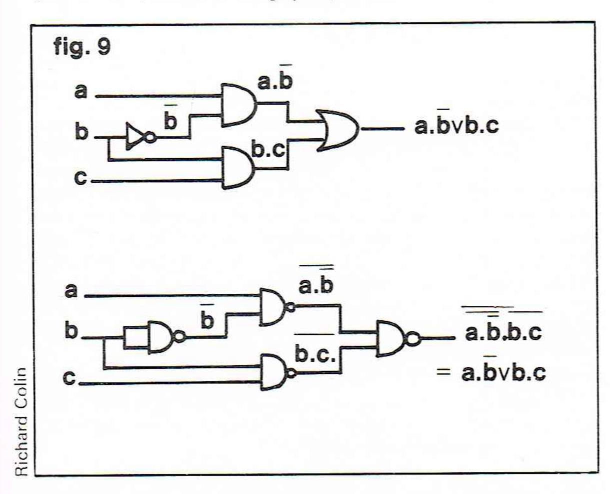
En logique négative, une porte NAND réalise une fonction NI et réciproquement (voir Théorème de Morgan) [*fig.* 7].

fig. 8

Circuit OU EXCLUSIF a, b → $a \oplus b$

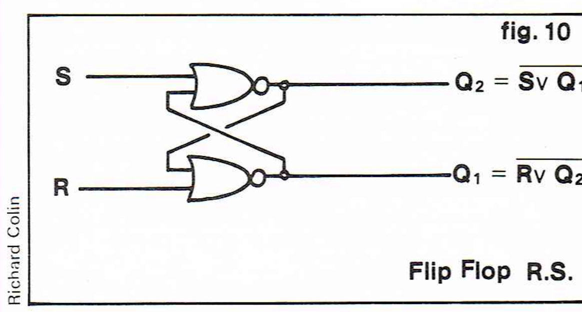
Richard Colin

Les schémas de la *figure* 9 représentent la fonction $S = a \cdot \bar{b} \vee b \cdot c$ en logique positive.



Une catégorie importante de circuits est constituée par les circuits séquentiels. Nous n'avions considéré que des fonctions logiques, dont les sorties ne dépendent que de l'état des entrées. Dans les circuits séquentiels, l'état des sorties dépend de l'état présent et des états précédents des entrées.

Considérons la *figure* 10 :



Supposons $S = R = 0$ et $Q_1 = 1$.

On a alors : $Q_2 = \overline{S \vee Q_1} = \overline{0 \vee 1} = 0$

d'autre part : $Q_1 = \overline{R \vee Q_2} = \overline{0 \vee 0} = 1$. Il n'y a pas contradiction avec l'hypothèse $Q_1 = 1$. Cet état est donc possible. Supposons toujours : $S = R = 0$ et $Q_1 = 0$

on a alors $Q_2 = \overline{S \vee Q_1} = \overline{0 \vee 0} = 1$

et $Q_1 = \overline{R \vee Q_2}$ $\overline{0 \vee 1} = 0$

Cet état est aussi possible. Ce circuit a donc deux états possibles quand $R = S = 0$. Supposons qu'il soit dans l'un de ces états, et que S devienne égal à 1.

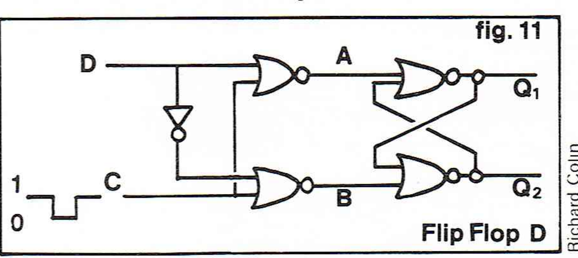
On aura $Q_2 = \overline{1 \vee Q_1} = 0$ quel que soit Q_1

et $Q_1 = \overline{R \vee Q_2} = \overline{0 \vee 0} = 1$

Si S redevient égal à 0, Q_2 est maintenu à 1 par Q_1 et Q_1 à 0 par Q_2. Cet état, étant stable pour $R = S = 0$, subsiste après que S soit redevenu égal à 0. Si on fait passer S plusieurs fois à 1 puis revenir à 0, l'état des sorties ne change pas, si R reste toujours à 0.

Par contre si c'est R qui passe à 1, puis revient à 0, c'est Q_2 qui prend la valeur 1 et Q_1 la valeur 0. Ce circuit permet donc de transformer une information passagère (passage temporaire à 1 puis retour à 0) en une information permanente. Il réalise donc une fonction de mise en mémoire. Selon l'état des sorties, on peut savoir laquelle des deux entrées est passée à 1 la dernière. Remarquons que si les deux entrées sont simultanément à 1, on aura $Q_1 = Q_2 = 0$. Lorsque les entrées reviendront à 0, c'est la dernière à revenir à 0 qui déterminera l'état du système. Un tel circuit est appelé Flip Flop R. S.

Considérons maintenant la *figure* 11 :



L'entrée C est appelée *entrée d'horloge* ou *entrée d'écriture*. Par rapport au Flip Flop R. S., nous avons ajouté un inverseur et deux portes. En l'absence de l'impulsion d'horloge ($C = 1$), on a $A = B = 0$; on a donc soit $Q_1 = 0$, $Q_2 = 1$, soit $Q_1 = 1$, $Q_2 = 0$. Pendant l'impulsion d'horloge ($C = 0$), on a $A = \bar{D}$ et $B = D$, donc un des points A et B est à 0, l'autre à 1. Les sorties Q_1 Q_2 vont prendre l'état correspondant, à savoir $Q_1 = 1$, $Q_2 = 0$ si $D = 1$ et $Q_1 = 0$, $Q_2 = 1$ si $D = 0$ pendant l'impulsion d'horloge.

Ce circuit permet donc de mémoriser l'état de l'entrée D au moment de l'impulsion d'horloge seulement.

Un tel circuit est appelé Flip Flop D. C'est un circuit dit de type synchrone, car il ne peut changer d'état qu'au moment de l'impulsion d'horloge, contrairement au Flip Flop R. S. (appelé asynchrone).

Structure et fonctionnement des ordinateurs

Arithmétique binaire dans les calculateurs

L'addition en binaire peut se réaliser simplement, chiffre à chiffre, comme en décimal, en utilisant la table d'addition :

$0 + 0 = 0$
$0 + 1 = 1$
$1 + 0 = 1$
$1 + 1 = 10$ Soit 0 et un report égal à 1.

Exemple

```
    (1) (1)
     1   1   1   0   0
+
     1   0   1   0   1
 1   1   0   0   0   1
```

Nous voyons que la table d'addition en binaire présente une grande ressemblance avec une table de vérité de fonction logique. On peut donc convenir de représenter un chiffre binaire par une variable logique. Pour simplifier les conventions on représentera le 1 de la numération à base deux par un 1 logique (le contraire serait possible). Dans ces conditions, l'addition consiste à déterminer chaque chiffre du résultat et la retenue en fonction des chiffres de même rang des opérandes et du report du rang précédent. On peut donc définir deux fonctions logiques : la somme modulo 2, S et la retenue R, de trois variables *a*, *b*, R' ; *a* et *b* représentent les chiffres des nombres à additionner, R' la retenue précédente.

◀ ***Figures 6, 7, 8 :* *représentation de circuits simples les plus courants correspondant à des fonctions réalisées en logique positive.* *Figure 9 :* *exemple de réalisation d'une fonction logique à partir de fonctions élémentaires.***

◀ ***Figure 11 ;* *circuit séquentiel de type synchrone : Flip Flop D.***

◀ ***Figure 10 ;* *circuit séquentiel de type asynchrone : Flip Flop R. S.***

a	*b*	R'	R	S
0	0	0	0	0
0	0	1	0	1
0	1	0	0	1
0	1	1	1	0
1	0	0	0	1
1	0	1	1	0
1	1	0	1	0
1	1	1	1	1

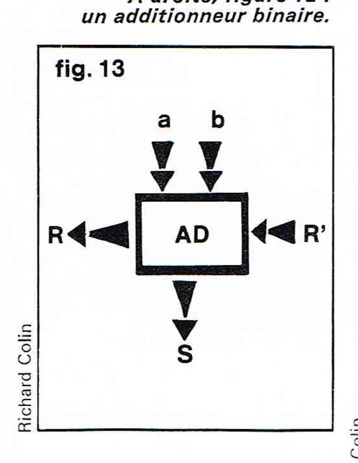
Les fonctions peuvent s'écrire

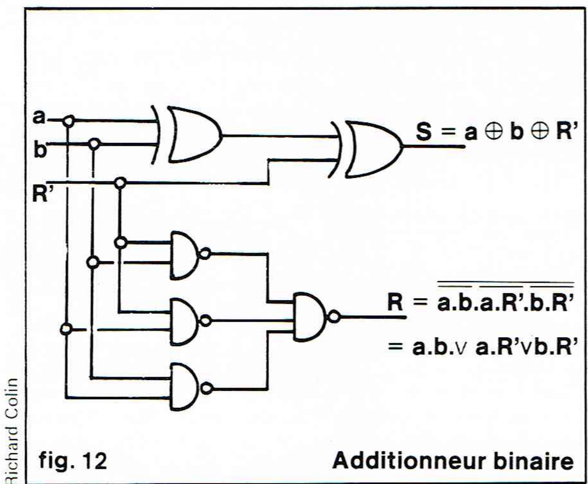
$$S = a \oplus b \oplus R'$$
$$R = a \cdot b \vee b \cdot R \vee a \cdot R$$

On pourra, par exemple, réaliser ces fonctions à l'aide de circuits logiques suivant la *figure* 12.

Nous pouvons enfermer ce circuit dans une « boîte noire » et le représenter désormais par la *figure* 13.

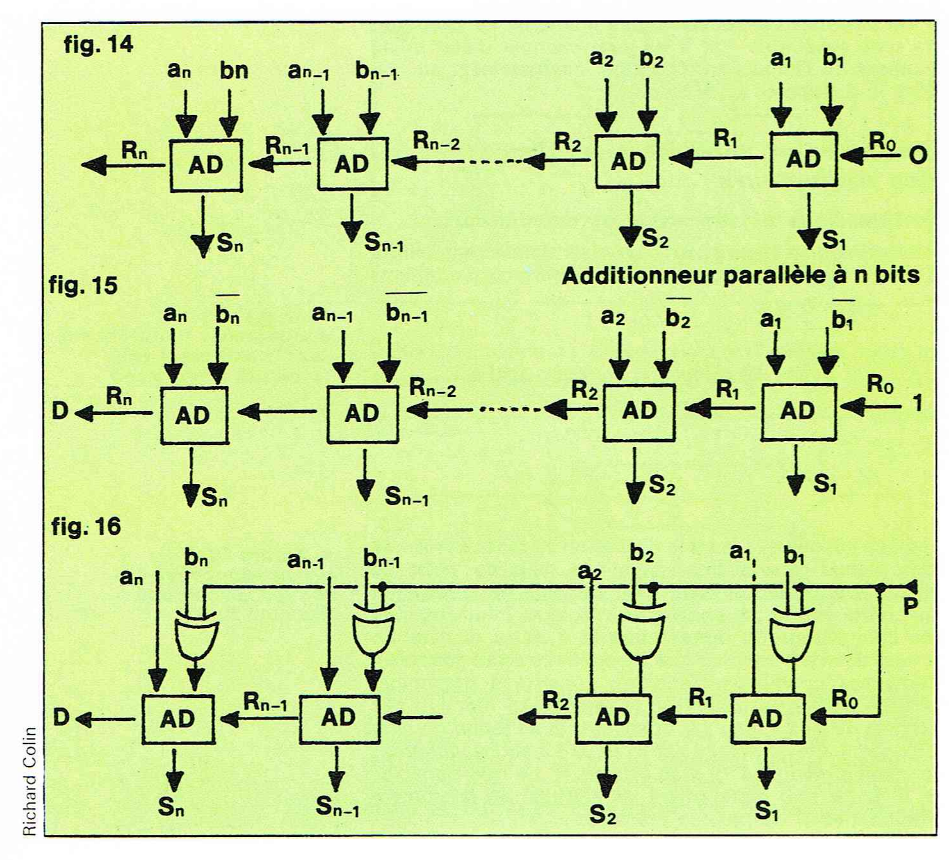
▼ A gauche, figure 13 : circuit de la figure 12 enfermé dans une « boîte noire ». A droite, figure 12 : un additionneur binaire.





En mettant en parallèle plusieurs circuits identiques, on pourra effectuer une addition, chaque report étant envoyé sur l'étage suivant (*fig.* 14).

▼ Figure 14 : additionneur parallèle à n bits. Figure 15 : représentation de la soustraction A — B. Figure 16 : additionneur soustracteur parallèle.



Si on utilise une représentation des nombres sur *n* bits, il se peut que le résultat réel s'exprime sur $n+1$ bits (Report $R_n = 1$). Dans ce cas il se produit un dépassement de capacité, car le bit supplémentaire est perdu. En général la retenue R_n est mémorisée, ce qui permet après l'opération de vérifier s'il y a eu ou non dépassement. Si on dispose de *n* bits on peut représenter naturellement les nombres de 0 à $2^n - 1$. On voit que si on ajoute 1 à $2^n - 1$, on obtiendra, avec notre additionneur, 0 comme résultat et un dépassement de capacité. L'additionneur travaille donc modulo 2^n. On peut utiliser cela pour traiter des nombres négatifs. En effet, soit *p*, $0 \leqslant p \leqslant 2^{n-1}$, on a : $-p \equiv 2^n - p$ (Mod. 2^n) ; on peut donc représenter $-p$ par $2^n - p$ qui est positif. Pour qu'une configuration binaire ne puisse pas représenter deux nombres différents, on représente de cette façon les nombres de -2^{n-1} à $2^{n-1} - 1$. Soit $0 \leqslant p < 2^{n-1}$, *p* contient au maximum $n - 1$ bits significatifs, donc le premier bit est 0 ; $-p$ est représenté par $p' = 2^n - p$, donc $2^{n-1} \leqslant p' \leqslant 2^n - 1$, donc le premier bit vaut 1. Le premier bit permet donc de connaître le signe d'un nombre. Cette représentation appelée « complément à 2 » est très souvent utilisée dans les calculateurs électroniques. Pour obtenir le complément à 2 de *p*, on complémente chaque bit et on ajoute 1 au résultat.

Exemple : soit $n = 5$
$p = 3$ en binaire $p = 00011$
on complémente les bits : 11100
on ajoute 1 : 11101 $= -p$

Vérification :

$$\begin{array}{r} 00011 \\ +\ 11101 \\ \hline (1)\ 00000 \end{array}$$

Le dernier report étant le dépassement, le résultat sur 5 bits est bien 0.

En représentant ainsi les nombres pour faire l'opération $p - q$ (ou $p + (-q)$), on fait en réalité l'opération $p + 2^n - q$, soit $2^n + p - q$. Si $p - q > 0$, le résultat est plus grand que 2^n, il y a dépassement et il reste $p - q$ qui est le résultat cherché. Si $p - q < 0$, le résultat est inférieur à 2^n, il n'y a pas de dépassement. On peut écrire $2^n + p - q = 2^n - (q - p)$. On obtient donc un résultat représentant un nombre négatif, puisque dans ce cas $q - p > 0$. On dispose ainsi d'un moyen pour faire des additions algébriques et des soustractions. Pour soustraire un nombre, on additionne son complément à 2.

On peut représenter la soustraction A — B par la *fig.* 15.

On applique sur les entrées les niveaux logiques *a* et $\bar{b}$ et 1 sur l'entrée de report du 1er étage, ce qui revient à former le complément à deux de B.

Si on veut pouvoir réaliser au choix une addition ou une soustraction, il faut avoir la possiblité de complémenter ou non les entrées *b* et d'appliquer un niveau 0 ou 1 sur l'entrée de report. La complémentation peut se faire à l'aide d'un circuit OU EXCLUSIF d'après la *figure* 16.

Si P vaut 0, les bits *b* ne sont pas complémentés et R_0 vaut 0, on fait donc l'opération A + B. Si P vaut 1, les bits *b* sont complémentés et R_0 vaut 1, l'opération est donc A — B.

Nous pouvons donc réaliser ces opérations. Cependant il faut que les opérandes soient présents sur les entrées et il faut prendre en compte le résultat. Pour cela, on utilise des registres. Un registre est un ensemble de flips flops permettant de mémoriser en même temps plusieurs bits. Le nombre et le rôle exact des registres varient d'un type de calculateur à l'autre. Cependant, on trouve très souvent la structure de la *figure* 17.



Un registre (RD) contient un opérande. Un autre registre (AC) contient l'autre opérande et sert à stocker le résultat. En pratique, l'opération se déroule ainsi : RD et AC sont chargés avec les opérandes. Leurs sorties étant connectées sur les entrées du circuit arithmétique, le résultat apparaît sur les entrées de AC. Une impulsion d'écriture dans AC mémorise le résultat à la place du contenu précédent, qui est perdu. Cette structure permet d'additionner (ou soustraire) successivement plusieurs nombres, en ajoutant chaque fois un nouveau nombre au contenu de l'accumulateur (d'où le nom d'accumulateur).

Multiplication - Division

La multiplication se fait en binaire comme en décimal, en multipliant le multiplicande par chaque chiffre du multiplicateur, et en ajoutant chaque fois le résultat décalé d'une position vers la gauche. En binaire, l'opération est particulièrement simple puisqu'on n'a à multiplier que par 0 ou 1. L'opération se réduit donc à des additions et des décalages.

Exemple :

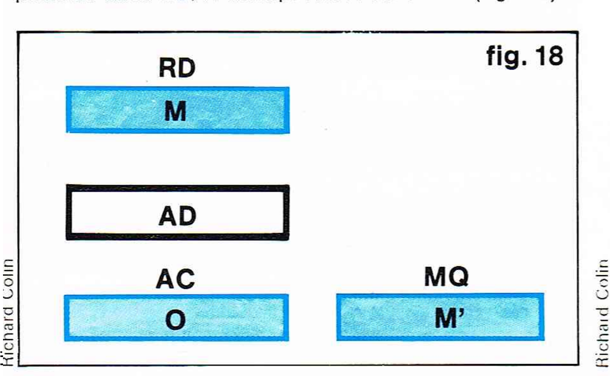
```
          1 0 0 1 0
        × 1 0 1 1 1
        -----------
          1 0 0 1 0
+       1 0 0 1 0
+     1 0 0 1 0
+   0 0 0 0 0
+ 1 0 0 1 0
  -----------------
= 1 1 0 0 1 1 1 1 0
```

Si les nombres sont exprimés sur n bits, le résultat peut avoir 2 n bits. En effet :

$$(2^n - 1) \times (2^n - 1) = 2^{2n} - 2^{n+1} + 1$$

La multiplication demande donc une capacité doublée. Pour cela, on associe à l'accumulateur un deuxième registre, appelé souvent MQ (Multiplicateur Quotient). On peut remarquer, cependant, que chaque addition ne porte que sur n bits, les bits de poids inférieur n'intervenant plus, l'un après l'autre. On peut alors réaliser la multiplication d'après l'algorithme qui suit.

Initialisation : on met AC à 0. On charge le multiplicande dans RD, le multiplicateur dans MQ (*fig.* 18).



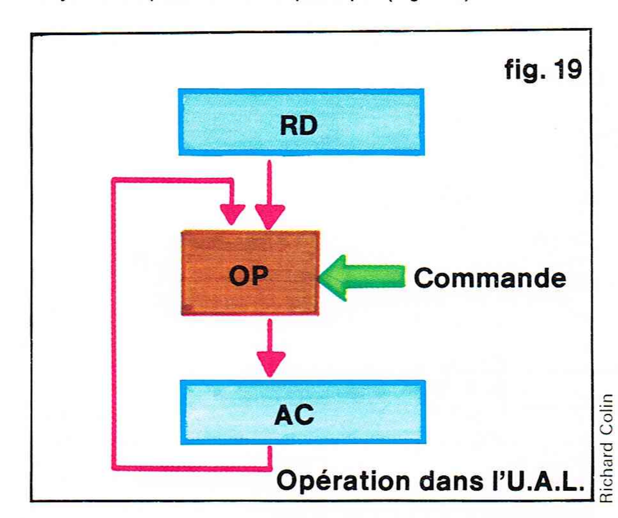
1 — Si le bit le plus à droite de MQ vaut 1, passer à l'étape 2. Sinon passer à l'étape 3.
2 — Additionner le multiplicande au contenu de AC.
3 — Décaler le contenu de l'ensemble AC — MQ d'une position vers la droite; le dernier bit de MQ est perdu. Décaler le résultat de l'addition à droite est équivalent à décaler le multiplicande à gauche lors de l'addition suivante.
4 — Recommencer n fois l'opération à partir de 1. Après n opérations le résultat se trouve dans l'ensemble AC — MQ.

On peut en déduire une méthode pour effectuer les divisions par soustractions successives, similaire à la division en décimal.

Initialisation : charger le dividende dans AC — MQ, le diviseur dans RD.
1 — Soustraire le contenu de RD (diviseur) du contenu de AC (poids forts du dividende).
Si le résultat dans AC est négatif, passer à l'étape 2.
Si le résultat est positif, passer en 3.
2 — Le dividende est plus grand que les poids forts du diviseur. Le chiffre obtenu pour le quotient est 0. Restaurer la valeur initiale de AC en ajoutant le dividende, puis passer à l'étape 4.
3 — Le chiffre obtenu pour le quotient est 1.
4 — Décaler le contenu de l'ensemble AC — MQ d'une position à gauche, en rentrant à droite de MQ la valeur trouvée pour le chiffre du quotient.
5 — Recommencer n fois l'opération à partir de 1. Après n opérations, le quotient se trouve dans MQ et le reste dans AC.

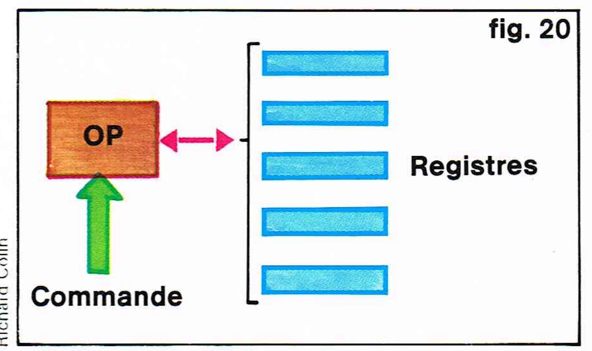
Unité arithmétique et logique (U. A. L.)

Avec un additionneur et des registres, il est possible de réaliser les quatre opérations arithmétiques. Il est facile d'imaginer effectuer d'autres opérations entre registres, toujours d'après le même principe (*fig.* 19).



◀ *Figure 19 :* ***OP désigne ici un circuit qui réalise l'opération spécifiée par les signaux de commande.***

Cet ensemble constitue l'unité arithmétique et logique. En fait, on peut en donner une définition plus générale comme étant un ensemble de circuits réalisant des opérations à caractère arithmétique ou logique entre les contenus d'un certain nombre de registres (*fig.* 20). Certaines opérations peuvent ne concerner qu'un seul registre.



◀ ***A gauche, figure 18 : schéma de l'initialisation pour réaliser la multiplication. A droite, figure 20 : opération dans l'U. A. L. entre les contenus d'un certain nombre de registres.***

Mémoire

L'un des points importants pour le traitement automatique est la nécessité pour l'organe de traitement de pouvoir stocker des informations. Il faut distinguer entre deux sortes de mise en mémoire.

— Stockage d'information en vue d'une utilisation ultérieure (par exemple des données ou des résultats d'un calcul). Dans ce cas il s'agit de produire ces informations sous une forme facile à réutiliser si besoin est.

— Stockage des informations immédiatement nécessaires à la conduite du traitement et à l'obtention des résultats. Dans ce cas, il faut pouvoir accéder rapidement à n'importe quelle information. Pour cela les calculateurs sont munis d'une mémoire centrale. La mémoire centrale est constituée d'un très grand nombre de registres, appelés aussi « mots mémoire », destinés à mémoriser les données d'un problème et les informations relatives au déroulement du traitement (le programme).

A mesure que le traitement s'effectue, les informations nécessaires sont transférées de la mémoire dans des

◀ ***Page ci-contre, figure 17 : structure permettant la réalisation d'opérations dans un accumulateur.***

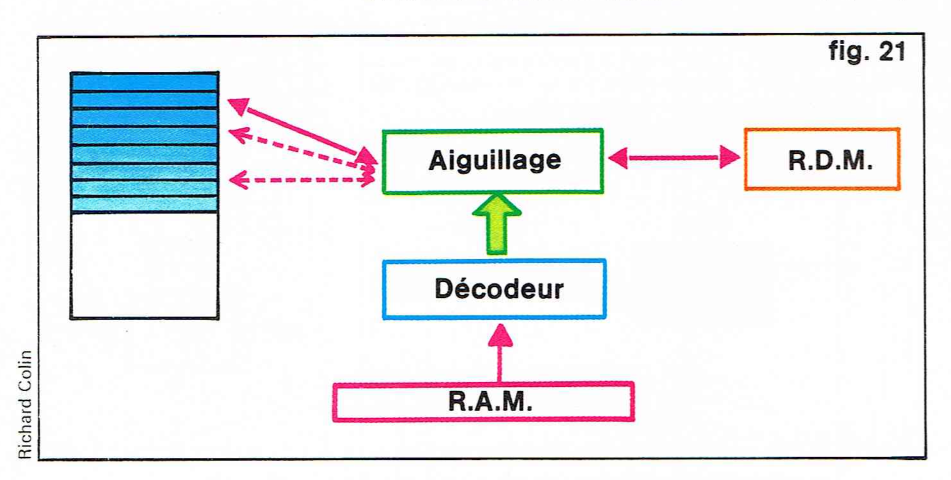
registres de l'U. A. L. ou d'autres registres, pour y être traitées. Des résultats peuvent être transférés de l'U. A. L. vers la mémoire.

Une information en mémoire est repérée par le numéro du registre qui la contient. C'est ce qu'on appelle l'adresse du mot mémoire.

A la mémoire sont associés deux registres particuliers.

— Le registre d'adresse mémoire (R.A.M.). Lorsqu'on veut accéder à un mot mémoire, son adresse est chargée dans le R.A.M. Un circuit décodeur d'adresse met alors le mot sélectionné en communication avec le registre de données mémoire.

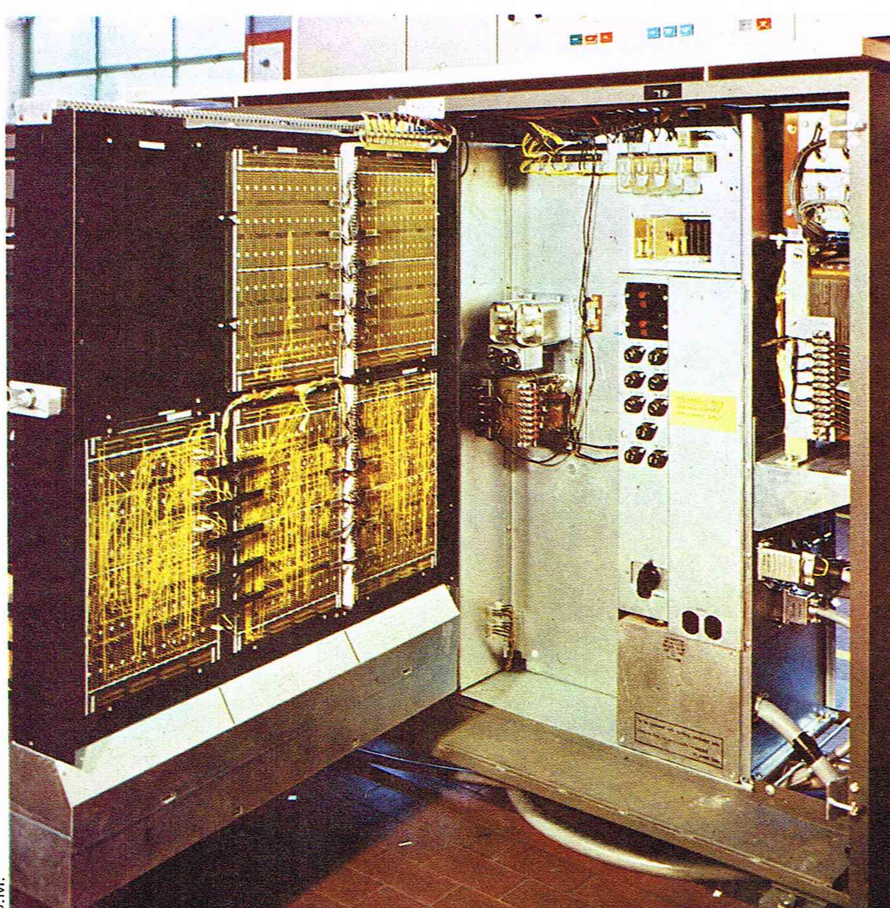
— Le registre de données mémoire. C'est par ce registre que transitent toutes les informations échangées entre la mémoire et d'autres organes du calculateur (*fig.* 21).

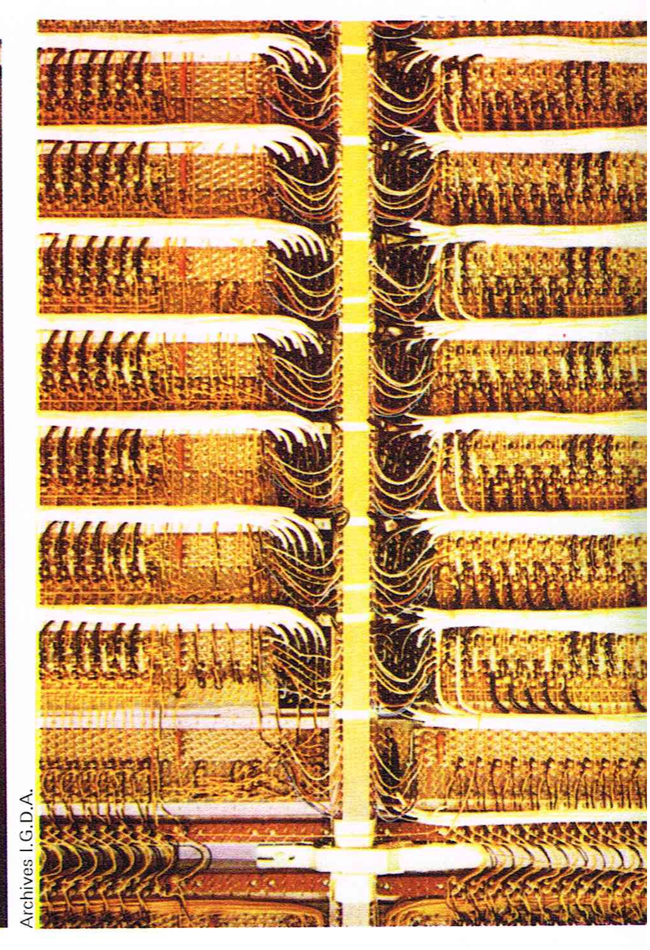


▲ *Figure 21 : représentation schématique de l'organisation de la* mémoire.

▶ *Page ci-contre, à gauche, figure 22 : structure de l'*unité de contrôle.

▼ *A gauche, partie interne de l'*unité centrale *d'un système 360/IBM. A droite, partie interne d'un calculateur électronique.*





Certains calculateurs ont une partie de leur mémoire constituée de flips flops électroniques, ce qui permet des échanges très rapides. Cependant, dans la grande majorité des cas, la mémoire est constituée de tores de ferrite. La ferrite présente un cycle d'hystérésis presque rectangulaire. Lorsqu'elle est excitée par un courant électrique circulant dans un circuit primaire, elle prend une aimantation dont le sens est lié au sens du courant. L'aimantation subsiste après l'arrêt du courant; si on force l'aimantation à changer de sens, on peut recueillir une impulsion électrique dans un circuit secondaire. On peut donc associer un sens d'aimantation à 0 et l'autre à 1. On écrit donc un 0 ou un 1 selon le sens du courant d'excitation. En lecture, on force l'aimantation dans le sens correspondant à 0. S'il y a changement de sens, c'est-à-dire si l'information était 1, on obtient une impulsion électrique au circuit secondaire (circuit de lecture).

Un tel procédé de lecture est destructif puisqu'on efface l'information pour la lire. A la lecture, l'information est transférée dans R.D.M. Les circuits sont conçus de façon à effectuer automatiquement une réécriture de l'information dans le mot qui vient d'être lu; ainsi l'information reste en mémoire.

Les mémoires à ferrite constituent des mémoires permanentes; elles gardent l'information même si l'alimentation électrique du calculateur est coupée, contrairement aux registres électroniques qui perdent l'information dans ce cas.

Unité de contrôle - Instructions machine

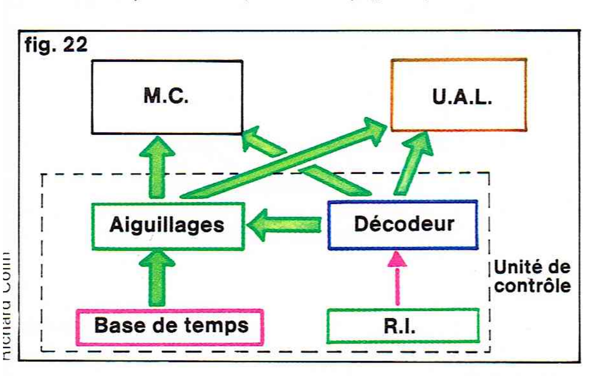
Nous avons vu le principe selon lequel les opérations sont effectuées dans l'U.A.L., et les échanges avec la mémoire. Ces opérations nécessitent certains signaux de commande qui peuvent être soit des niveaux logiques appliqués sur certaines entrées, destinées à déterminer l'opération à effectuer, soit des impulsions destinées à commander successivement les différentes phases de l'opération. Ces signaux sont délivrés par l'unité de contrôle, qui comporte essentiellement deux éléments :

— Un registre d'instructions (R.I.) destiné à recevoir les instructions, associé à un décodeur d'instruction qui, en fonction du code contenu dans R.I., génère les

niveaux logiques de commande et d'aiguillage des impulsions d'horloge.

— Une base de temps qui génère en permanence toutes les impulsions qui peuvent être nécessaires. Ces impulsions sont aiguillées vers les circuits correspondants par le décodeur d'instructions.

Chaque instruction correspond à une opération élémentaire réalisable par les circuits des différentes unités. Ces instructions sont évidemment codées en binaire. Ce code constitue le langage machine. C'est le seul que l'unité de contrôle peut « comprendre » (*fig.* 22).



Généralement, une instruction comprend plusieurs parties qui sont décodées séparément :

— Une partie « Code Opération » qui indique le type d'opérations à effectuer.

— Une partie « opérandes » qui spécifie où et comment les opérandes doivent être pris. Selon les opérations, les opérandes peuvent être soit en mémoire, soit dans l'U.A.L., soit dans l'instruction elle-même. Cette partie est donc interprétée en fonction du code opération. *A priori*, le nombre d'opérandes dépend de l'opération. Par exemple, pour une division, on pourrait faire figurer dans l'instruction les adresses où prendre le diviseur, le dividende, et où ranger le quotient et le reste. Cela ferait quatre adresses à faire figurer dans l'instruction. D'autre part, les instructions d'un programme étant stockées en mémoire, on pourrait également faire figurer l'adresse de l'instruction suivante à exécuter.

En fait, dans la plupart des cas, on ne fait figurer qu'une adresse par instruction machine, en décomposant certaines opérations en plusieurs étapes de nature différente :

— transferts d'un seul opérande à la fois entre mémoire et U.A.L. ;

— opérations portant sur un seul opérande pris en mémoire, les autres étant pris dans l'U.A.L.

Ainsi une addition peut se faire en trois instructions :

— charger l'accumulateur avec le premier opérande ;

— additionner le deuxième opérande au contenu de l'accumulateur ;

— ranger le contenu de l'accumulateur (résultat) en mémoire.

Une telle décomposition est avantageuse pour les calculs enchaînés. On n'a pas toujours besoin de ranger en mémoire des résultats intermédiaires, ils restent dans l'accumulateur (ou dans d'autres registres de l'U.A.L.).

Pour la succession des instructions, on range simplement les instructions successives à des adresses consécutives. Dans certains cas, après une opération, il faut effectuer une séquence d'instructions qui n'est pas rangée immédiatement à la suite. Cela peut se faire par une instruction de branchement, qui n'effectue aucune opération, mais qui indique à l'unité de contrôle l'adresse où commence la séquence suivante. Cela permet les ruptures de séquence, notion fondamentale pour l'utilisation des ordinateurs.

Les ruptures de séquences

Il est souvent nécessaire de prendre des décisions en cours de traitement. En fonction de certains résultats, la suite des calculs peut se dérouler de différentes façons (voir l'exemple d'équation du second degré en début d'article). La décision consiste à choisir la séquence suivante, au moyen d'instructions de branchement (*fig.* 23). Ces décisions ne sont possibles que si on peut effectivement accéder à n'importe laquelle des séquences possibles. On voit ici apparaître l'importance de la notion de programme enregistré. Si tout le programme est présent en mémoire, n'importe quelle séquence est accessible à tout moment. Au contraire, si les opérations étaient commandées extérieurement, cela ne serait pas possible, car cela supposerait un changement de l'organe de commande. Par exemple, pour un programmateur électromécanique, il faudrait modifier la disposition des contacts.

Cette notion de rupture de séquence offre plusieurs possibilités très précieuses :

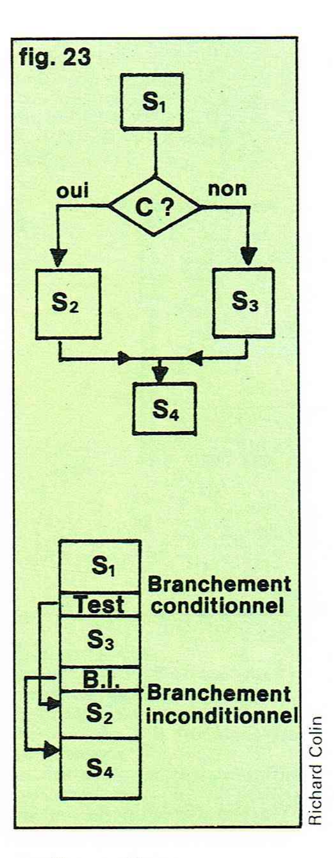
— Comme nous l'avons vu, on peut, par des **tests sur des résultats,** choisir la suite du traitement.

— **Boucles de programme.** Lorsqu'une même séquence doit être exécutée plusieurs fois de suite, on peut, en fin de séquence, faire un branchement au début. Pour que le programme ne boucle pas indéfiniment, il faut que la séquence comporte un test de sortie de la boucle (*fig.* 24).

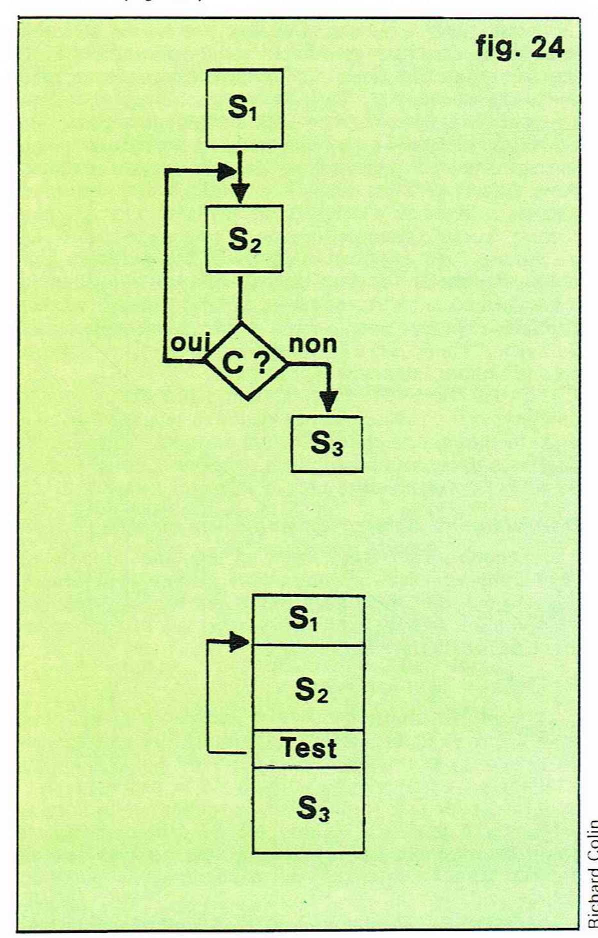
Par exemple, on peut compter les itérations et se brancher à une autre séquence après un nombre d'itérations déterminé.

On peut aussi tester un résultat et sortir de la boucle quand ce résultat répond à un certain critère ; on voit ici que le nombre d'itérations n'est pas fixé *a priori* dans le programme mais dépend des données du problème ; cela ne serait pas possible dans le cas d'une commande extérieure, et est donc lié à la notion de programme enregistré.

— **Sous-programmes.** Lorsqu'une séquence d'instructions doit être effectuée à différentes reprises au cours d'un traitement, on peut ne la faire figurer qu'une fois en mémoire. Lorsqu'on veut la faire exécuter, on y effectue un branchement en gardant en mémoire l'adresse d'où s'est fait le branchement. La séquence se termine par un branchement à l'adresse ainsi mémorisée. De cette façon, on peut « appeler » cette séquence nommée sous-programme de n'importe quel point du programme. Lorsqu'elle est terminée, l'exécution du programme appelant reprend là où elle s'était interrompue. Un sous-programme peut lui-même faire appel à d'autres sous-programmes. Cela permet, en outre, une conception modulaire du traitement, chaque sous-programme pouvant être écrit indépendamment, et utilisé par plusieurs programmes différents (*fig.* 25).



▲ ***Figure 23 :*** ***organigramme et implantation en mémoire d'un traitement avec branchement conditionnel. Exécuter la séquence*** S_1 ***;*** ***si la condition C est remplie, exécuter*** S_2 ***puis*** S_4 ***sinon*** S_3 ***puis*** S_4***. Les différentes séquences étant les unes derrière les autres en mémoire, il est nécessaire de terminer*** S_3 ***par un branchement en*** S_4***, pour ne pas exécuter*** S_2 ***après*** S_3***.***

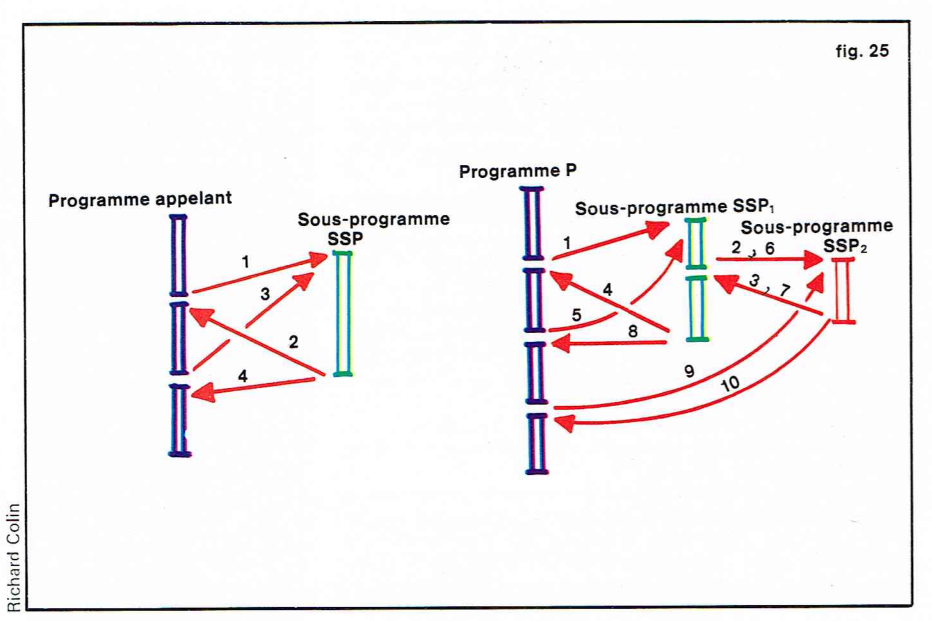


◀ ***Figure 24 :*** ***organigramme et implantation en mémoire d'une boucle.***

mémoire (R.A.M.).
— Un cycle de lecture est déclenché; le contenu du mot dont l'adresse est dans le R.A.M. est recopié dans le registre de données mémoire (R.D.M.).
— Le contenu de R.D.M. est recopié dans le registre d'instruction (R.I.). A ce moment, l'instruction est en place dans R.I.; elle peut être décodée et exécutée. Simultanément, le contenu de C.O. est incrémenté, c'est-à-dire que C.O. contient maintenant l'adresse suivante.

Exécution

Le code opération de l'instruction contenue dans R.I. est décodé par le décodeur d'instruction qui active les circuits correspondant à l'opération. Il faut alors distinguer entre les différents types d'opérations possibles.
— **Instructions comportant une adresse :**
• *avec lecture d'un opérande en mémoire* (*fig.* 26)
★ la partie adresse de l'instruction est recopiée dans R.A.M;
★ un cycle de lecture est déclenché;
★ le contenu de R.D.M. est pris en charge par les circuits préalablement activés par le décodeur d'instruction, et l'opération est exécutée.



▲ ***Figure 25 :*** *exemples de branchements à des sous-programmes. Les traits doubles représentent des séquences d'instructions, les flèches des branchements, les numéros l'ordre des branchements.* Ex. 1 : *le programme* P *appelle deux fois le sous-programme* SSP. Ex. 2 : *le programme* P *appelle deux fois le sous-programme* SSP1 *et une fois* SSP2. *Le sous-programme* SSP1 *appelle* SSP2 *une fois, chaque fois qu'il est lui-même appelé.*

Comme nous l'avons vu dans ce qui précède, les branchements sont souvent liés à des tests. Les instructions de branchements conditionnels consistent en général à effectuer le branchement à l'adresse indiquée si une condition est remplie; sinon l'exécution se poursuit en séquence. Les conditions testées sont par exemple le signe du contenu de l'accumulateur, l'égalité à zéro, le dépassement de capacité.

Unité centrale

L'unité arithmétique et logique, l'unité de contrôle et la mémoire centrale constituent l'unité centrale. Les caractéristiques peuvent varier considérablement d'un modèle d'ordinateur à l'autre (taille des mots mémoire, nombre de mots en mémoire, nombre d'instructions élémentaires, temps d'exécution d'une instruction), mais tous possèdent cette structure générale. Certains points de structure relèvent de conceptions différentes. On peut citer le rôle des registres et le mode d'adressage. Les registres peuvent être spécialisés, c'est-à-dire que chaque registre correspond à une fonction bien définie; on peut rencontrer au contraire un certain nombre de registres généraux pouvant tous servir à un certain nombre de fonctions telles que : accumulateur, multiplicateur quotient, registre de base ou d'index (voir plus loin). On peut rencontrer aussi différents modes d'adressage selon les machines. Dans certains calculateurs, les adresses sont obtenues directement dans les instructions. Dans d'autres, les instructions ne comportent qu'une adresse relative (appelée déplacement), définie par rapport à une adresse de base. A l'exécution, les adresses réelles sont obtenues en additionnant déplacement et base.

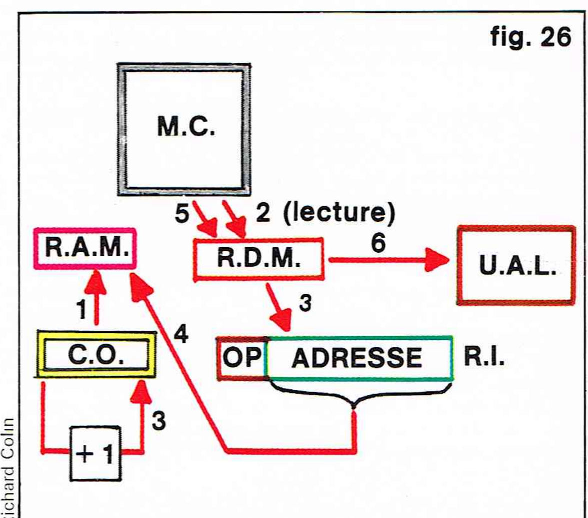
Tous ces choix sont généralement dictés par des considérations sur certaines caractéristiques telles que : taille de la mémoire, opérations élémentaires réalisables par les différents circuits opérationnels, de façon à constituer un ensemble cohérent et le plus performant possible.

Déroulement d'un programme en machine

L'exécution d'un programme en machine consiste en l'exécution successive d'opérations élémentaires suivant des séquences déterminées. Nous allons voir comment se déroulent l'exécution d'une instruction et l'enchaînement des instructions.

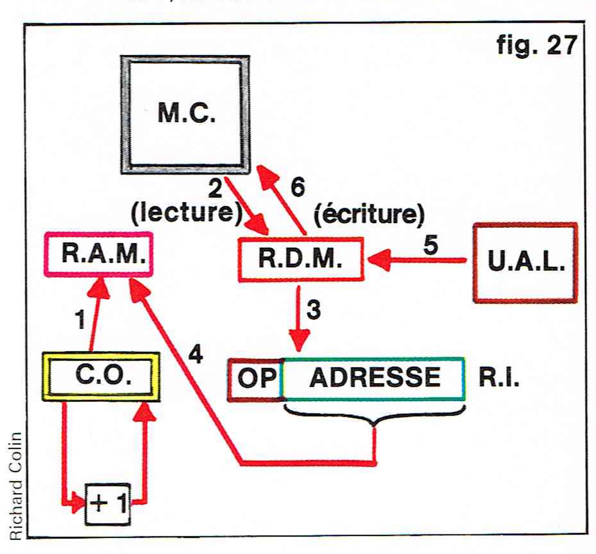
Recherche de l'instruction

La première phase consiste à accéder à l'instruction devant être exécutée. Il faut donc connaître son adresse en mémoire; pour cela, un registre de l'unité centrale contient en permanence l'adresse de la prochaine instruction à exécuter. Nous verrons comment cette adresse est tenue à jour. Ce registre est généralement appelé compteur ordinal; par la suite, il sera désigné par les initiales C.O. La recherche se décompose en plusieurs étapes :
— Transfert du contenu du C.O. dans le registre d'adresse



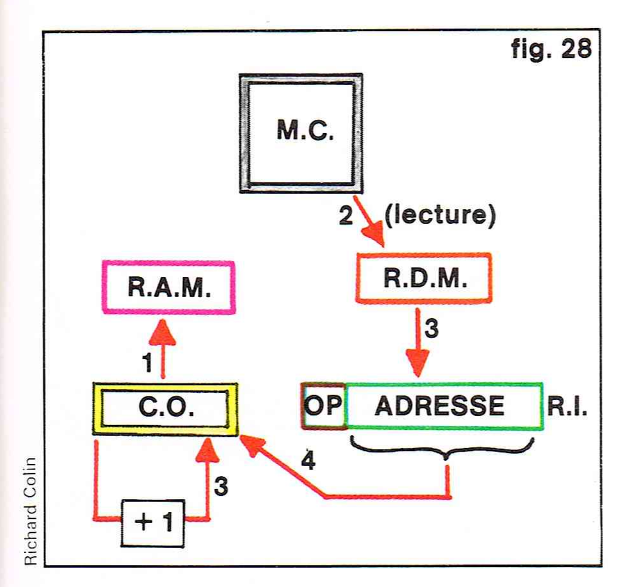
▶ ***Figure 26 :*** *transferts successifs entre les registres et l'unité centrale, lors de l'exécution d'une* instruction comportant une lecture en mémoire. *C.O. est incrémenté en même temps que l'instruction est chargée dans R.D.M.*

• *instruction d'écriture en mémoire* (*fig.* 27) :
★ la partie adresse de l'instruction est recopiée dans R.A.M;
★ le contenu du registre à ranger en mémoire est recopié dans R.D.M;
★ un cycle d'écriture est déclenché.



▶ ***Figure 27 :*** *instruction comportant une écriture en mémoire.*

• *l'instruction est une instruction de branchement* (conditionnel ou inconditionnel) : si les conditions de branchement sont réunies, ou s'il s'agit d'un branchement inconditionnel, la partie adresse de l'instruction est recopiée dans C.O (*fig.* 28).



— **Instructions ne comportant pas d'adresse**

Ce sont des opérations ne faisant pas intervenir la mémoire, mais uniquement des informations déjà présentes dans les registres de l'unité centrale. L'opération est alors exécutée conformément à l'instruction.

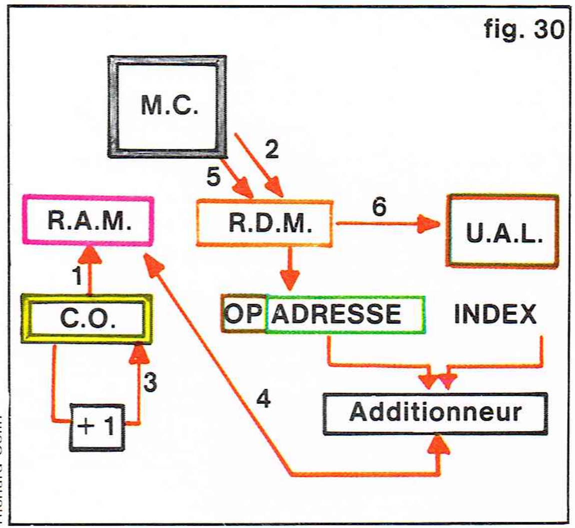
Dans tous les cas, le compteur ordinal contient l'adresse de l'instruction suivante à exécuter, le calculateur recommence alors les différentes étapes pour l'instruction suivante, et le programme se déroule ainsi. Il est seulement nécessaire de charger l'adresse de démarrage dans le compteur ordinal, avant de lancer l'exécution du programme; ensuite tout se déroule sans autre intervention. Le processus décrit ci-dessus n'est en fait que le processus de base. Il peut s'ajouter un certain nombre d'étapes intermédiaires. Celles-ci dépendent de la structure détaillée de l'unité centrale et du type d'instruction rencontré.

Adressage indirect

Au lieu de spécifier dans l'instruction l'adresse de l'opérande, on indique l'adresse d'un mot contenant l'adresse de l'opérande. Ce procédé est employé lorsqu'on ne peut pas connaître *a priori* l'adresse de l'opérande; il est utilisé, par exemple, pour les transferts d'information à des sous-programmes. La seule information qu'un sous-programme reçoit est l'adresse de retour, qui lui est transmise au moment de l'appel. Si les informations dont il a besoin ont été placées dans des mots ayant des positions définies par rapport à l'adresse de retour, le sous-programme peut calculer les adresses où trouver ses arguments et aller les chercher par adressage indirect. A l'exécution d'un adressage indirect, il y a un cycle mémoire supplémentaire, le contenu du mot adressé dans l'instruction étant transféré dans le R.I. à la place de l'adresse qui y figurait, avant la phase d'exécution (*fig.* 29).

Indexage

Certaines machines sont pourvues de registres d'index. Lorsqu'une instruction indexée est décodée, le contenu du registre d'index est ajouté à l'adresse figurant dans l'instruction. On peut ainsi, à partir d'une seule adresse, accéder à plusieurs mots mémoire en faisant varier le contenu du registre d'index. Ce procédé est très utile lorsqu'on a à traiter des informations contenues dans des tables (une table est constituée de plusieurs mots successifs contenant des informations de même type). On peut en effet explorer toute la table en n'utilisant que la première adresse (*fig.* 30).



D'après la façon dont se déroulent les opérations, nous pouvons voir qu'il n'y a aucune différence de nature entre les instructions et les données. Tout est en mémoire. Les informations représentant des instructions doivent seulement être structurées correctement, pour éviter que des données ne soient transférées dans le registre d'instructions et interprétées comme des instructions. Cela peut se produire, par exemple, lorsqu'en rédigeant un programme on fait une erreur dans un calcul d'adresse se traduisant par un branchement à une mauvaise adresse. De plus, il est possible, pour le programme, de déplacer ou modifier certaines de ses instructions. En effet, celles-ci étant en mémoire, peuvent être transférées dans l'U.A.L. pour y être traitées tout comme des données, puis remises en mémoire.

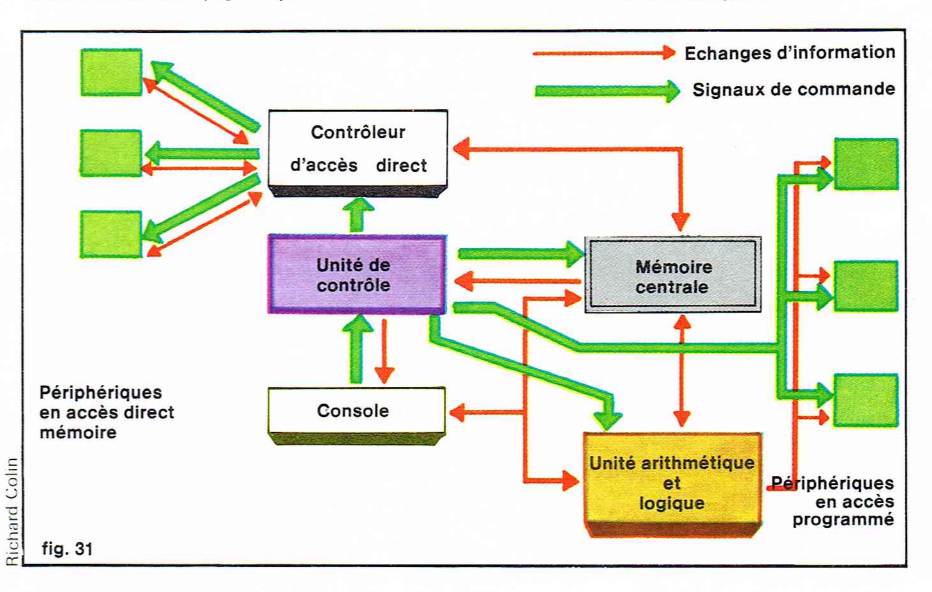
Les entrées-sorties

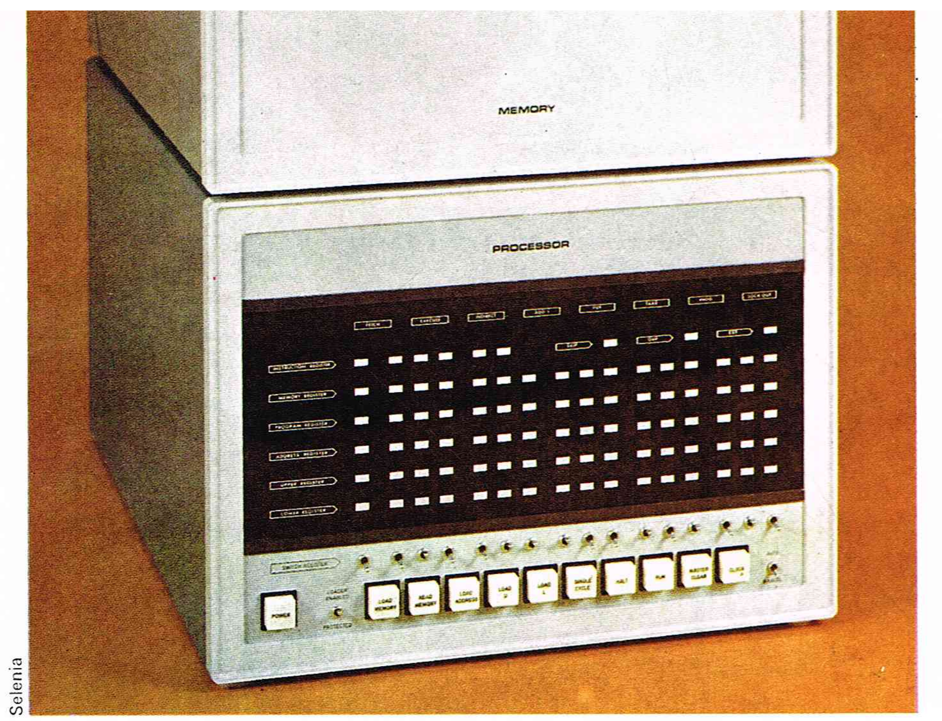
Le calculateur doit échanger des informations avec le monde extérieur (chargement des programmes, lecture des données, sortie de résultats). Pour cela il est muni d'unités d'entrées-sorties (*fig.* 31).

◀ Figure 28 : instruction de branchement; la partie adresse de l'instruction est transférée dans C.O. à la place du contenu précédent.

◀ Figure 30 : instruction avec adressage indexé.

▼ A gauche, figure 29 : instruction comportant une lecture avec adressage indirect. A droite, figure 31 : structure générale des calculateurs électroniques.





Selenia

▲ **Console** ***du calculateur Selenia GP-16.***

La console du calculateur

La console est directement associée à l'unité centrale. Elle permet de commander manuellement certaines opérations qui, ordinairement, se déroulent de façon interne. Elle comporte deux types d'éléments :

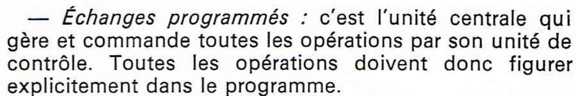
— Des interrupteurs (ou clés), poussoirs, etc., permettant de coder des informations binaires, de les charger en mémoire ou dans certains registres, de provoquer des opérations élémentaires en déclenchant certains cycles au niveau de l'unité de commande.

— Des voyants lumineux permettant de visualiser les contenus de mots mémoire, de registres et l'état de l'unité centrale.

Elle permet donc, en principe, d'utiliser le calculateur, en chargeant programmes et données manuellement et en regardant les résultats directement en mémoire. Il est évident qu'une telle méthode n'est pas très efficace. Elle est employée uniquement pour vérifier des points de détail dans le fonctionnement du calculateur ou dans le déroulement d'un programme.

Les périphériques

Les périphériques sont des organes d'entrées-sorties, plus ou moins spécialisés, permettant des communications plus rapides et sous des formes plus appropriées que la console du calculateur. Leur fonction consiste à transposer une représentation des informations interne au calculateur (donc binaire) en d'autres représentations, sur des supports physiques variés ou *vice versa*. Il existe différents types de périphériques, cependant les échanges avec l'unité centrale se font pratiquement toujours selon les mêmes principes. Ces échanges peuvent se faire de deux façons :



— *Échanges programmés :* c'est l'unité centrale qui gère et commande toutes les opérations par son unité de contrôle. Toutes les opérations doivent donc figurer explicitement dans le programme.

— *Échanges en accès direct :* certains périphériques sont munis de circuits de commandes indépendants, et peuvent accéder par eux-mêmes à la mémoire centrale. Dans ce cas, le rôle de l'unité centrale consiste à initialiser le contrôleur du périphérique qui prend en charge le détail des opérations. De cette façon, l'unité centrale peut effectuer un travail interne pendant que le périphérique effectue ses transferts.

En plus des informations lues ou écrites, l'U.C. et le périphérique échangent des informations de contrôle. L'U.C. envoie des commandes et peut tester l'état du périphérique. En effet, les périphériques, comportant presque toujours des organes mécaniques, sont beaucoup plus lents que l'U.C. Celle-ci doit donc attendre que le périphérique ait terminé une action avant d'envoyer une nouvelle commande.

Il existe plusieurs types de périphériques, correspondant à différents besoins.

Communication directe homme-machine

— *Clavier alphanumérique :* semblable à un clavier de machine à écrire, il transmet au calculateur les codes correspondant aux touches frappées.

— *Téléimprimante :* elle imprime sur papier les caractères d'imprimerie correspondant aux codes transmis par le calculateur.

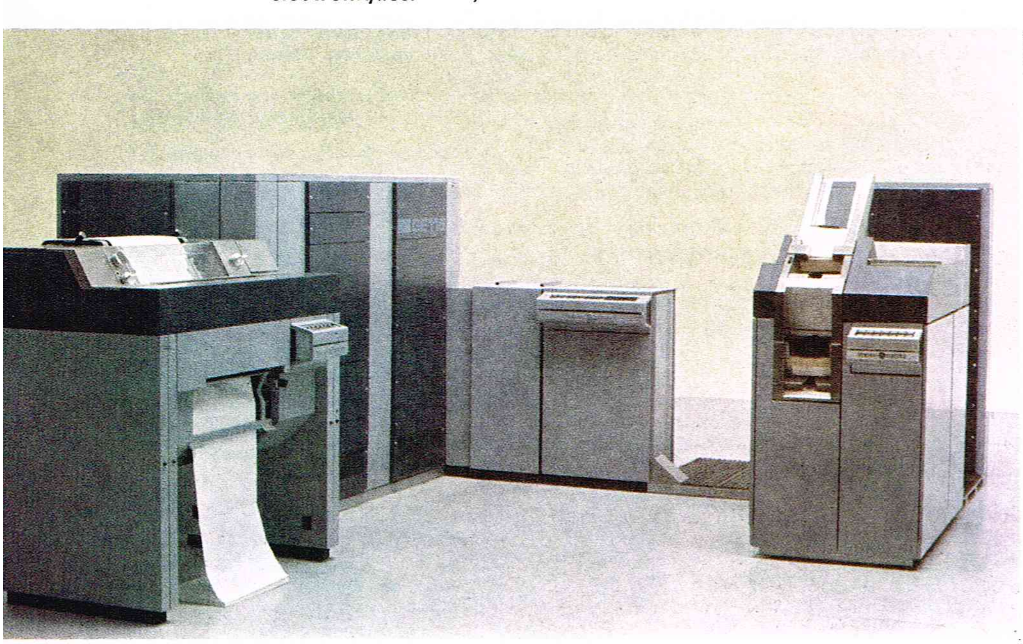
— *Écrans de visualisation à rayons cathodiques :* en affichant des points lumineux à des positions correspondant aux valeurs transmises par le calculateur, ils permettent de représenter soit des dessins, soit des caractères (donc des textes).

On trouve très souvent associés un clavier et une téléimprimante ou un écran. Cette combinaison permet le dialogue entre un opérateur, qui transmet des informations par le clavier, et un programme qui transmet des informations par l'écran ou la téléimprimante. Ce dialogue ne peut avoir lieu que sous la forme prévue dans le programme. Les programmes d'entrées-sorties comportent toujours des tests sur la nature de l'information qu'ils reçoivent. Par exemple, si l'opérateur doit entrer un nombre, et qu'il frappe un caractère non numérique (il arrive qu'on frappe la lettre O à la place du chiffre zéro), le programme ne doit pas en tenir compte, et recommencer la séquence de lecture.

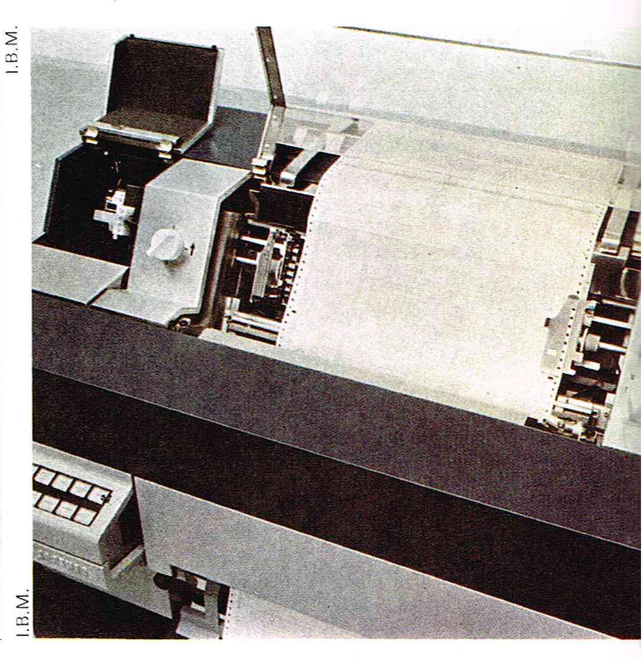
Entrées-sorties sur cartes ou rubans perforés

On utilise beaucoup les cartes ou les rubans perforés, lorsqu'un dialogue n'est pas nécessaire entre l'utilisateur et le calculateur. Cartes ou rubans peuvent être perforés sur une perforatrice indépendante du calculateur, et lus en bloc par un lecteur connecté au calculateur.

▼ *A gauche, système GE-115 à cartes (General Electric) :* **la carte perforée** ***représente l'unique support pour l'introduction, la mémoire et le traitement des données de ce système. A droite, détail d'une*** **imprimante** ***reliée aux systèmes électroniques.***



I.B.M.



I.B.M.

On peut éventuellement faire perforer des résultats si on veut les réutiliser comme données pour d'autres programmes.

Imprimantes rapides

Très souvent, les résultats d'un programme doivent simplement être imprimés. Les imprimantes rapides impriment une ligne à la fois, contrairement aux téléimprimantes, associées à un clavier, qui impriment caractère par caractère.

Traceurs de courbes

Ils permettent de dessiner des figures, en commandant les déplacements d'une plume sur un papier.

Mémoires de masse

Pour stocker de grandes quantités d'informations on utilise des unités :

à bande magnétique,
à disque magnétique,
à tambour magnétique,
à feuillets magnétiques.

Ces périphériques ne sont pas destinés à la communication directe ou indirecte homme-machine. Ils permettent le stockage d'informations avec les avantages suivants :

— La même unité permet lecture et écriture sur le même support.

— Ils permettent d'accéder à toute l'information qu'ils contiennent, quel que soit l'ordre des opérations de lecture ou écriture qu'on veut effectuer.

— Ils ont des capacités élevées.

— Ils présentent des vitesses de transfert avec l'U.C. très élevées.

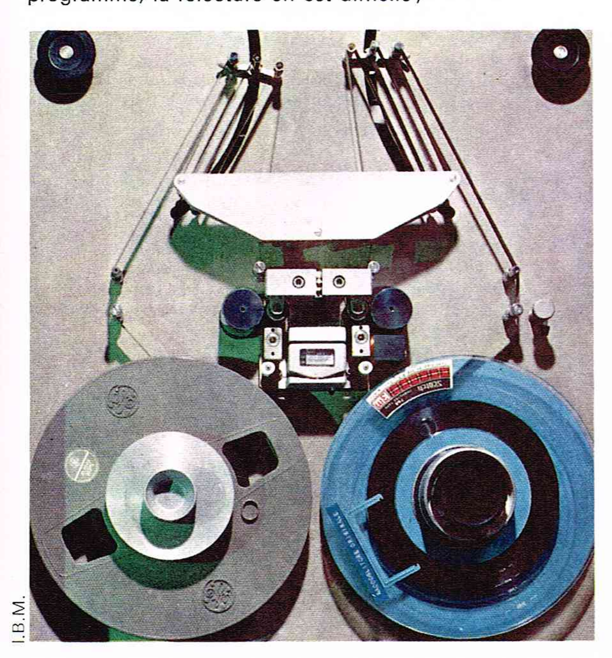
Périphériques spéciaux

Des équipements automatiques de mesure de grandeurs physiques (pressions, températures, tensions électriques, etc.) peuvent être connectés à un calculateur. Ainsi le calculateur peut enregistrer et surveiller ce qui se passe, par exemple dans une installation industrielle ou expérimentale. De même le calculateur peut être connecté à des organes de télécommande et commander lui-même ces installations.

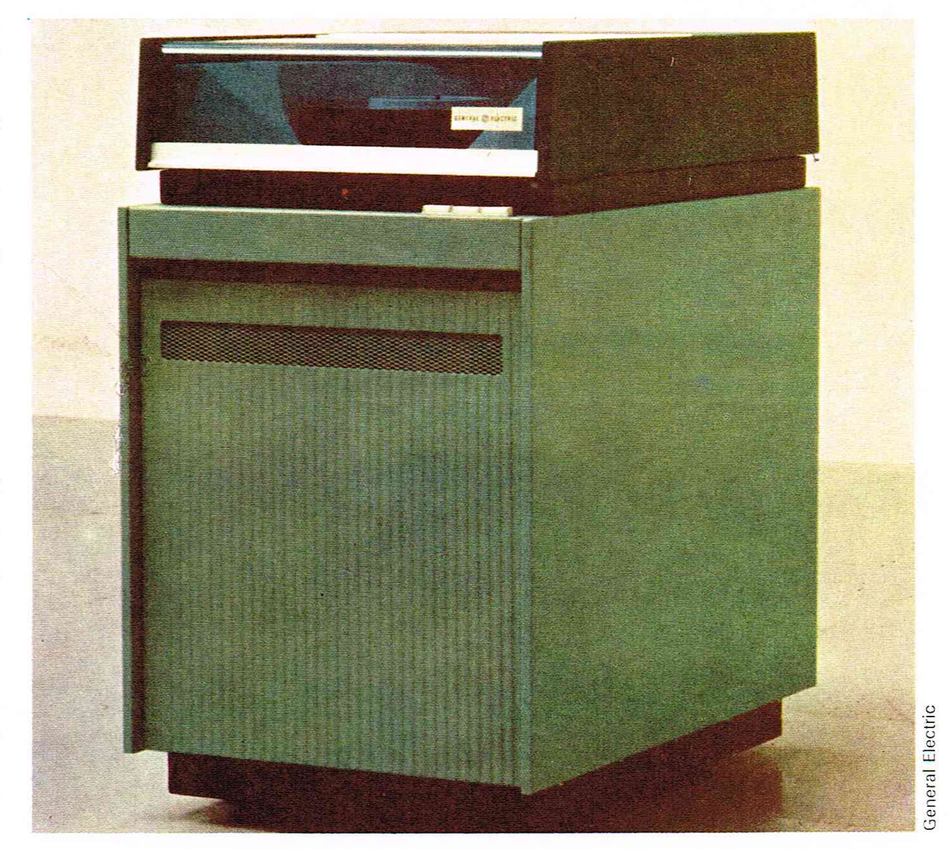
Écriture et manipulation des programmes

Dans ce qui précède nous avons toujours considéré des cas d'exécution de programmes supposés en mémoire, sans préciser la manière de les y faire parvenir. Le programme, tel qu'il se présente en mémoire lors de l'exécution, est sous forme binaire, en langage machine. On peut envisager d'écrire ce programme directement en binaire, et de le charger en mémoire à l'aide des clés de la console. C'est, apparemment, la méthode la plus directe, elle présente cependant de graves inconvénients :

— les risques d'erreurs sont importants à l'écriture du programme, la relecture en est difficile ;



I.B.M.



General Electric

▲ ***Unité à* disques magnétiques ; *le disque présente par rapport au ruban le grand avantage de permettre un traitement non séquentiel des données.***

— l'introduction manuelle présente d'autant plus de risques de fausses manœuvres que le programme est plus long ; l'opération est lente et immobilise le calculateur pendant toute sa durée.

L'emploi de cette méthode ne peut donc pas être généralisé, cependant nous verrons quelques cas où cela peut être utile. Nous voyons que l'utilisation pratique d'un calculateur pose deux types de problème : — l'écriture des programmes, et — la manipulation des programmes (chargement). Pour faciliter ces opérations, on a recours à un ensemble de programmes spécialisés ; nous allons voir comment.

L'écriture des programmes

Le nombre des opérations élémentaires que peut réaliser un calculateur, peut être plus ou moins grand, selon le type de machine, mais c'est toujours un nombre fini, allant de quelques instructions à quelques dizaines en ce qui concerne les instructions courantes. Il est alors plus pratique d'utiliser des symboles, plus proches de la signification de l'opération, que des codes purement numériques, sans rapport avec l'opération et donc plus difficiles à retenir et à ne pas confondre. Par exemple, on peut convenir de représenter l'addition par le symbole ADD, la division par DIV, le branchement inconditionnel par JMP (en anglais *jump* = saut), etc. De même, on peut utiliser des symboles pour repérer des informations (nombres ou instructions) à l'intérieur même du programme, par exemple appeler X, Y, ou Z (ou de toute autre façon) des adresses dans le programme auxquelles on a besoin de faire référence. Ceci est à rapprocher de l'utilisation, en mathématiques, de lettres pour représenter les nombres, avec cependant la distinction importante qu'en informatique on représente les adresses et non les contenus.

Ainsi, l'instruction ADD X signifie : ajouter au contenu actuel de l'accumulateur le contenu du mot mémoire d'adresse X, et non la valeur X qui, elle, est une adresse. L'avantage d'une telle notation est qu'elle permet de faire référence, lors de l'écriture d'un programme, à des informations sans avoir besoin de connaître explicitement leur adresse. La valeur exacte de l'adresse ne sera en effet déterminée que plus tard, mais dans tout le programme, un même symbole représente toujours une seule adresse (ou une seule opération s'il s'agit d'un code opération).

◄ ***Détail d'une* mémoire à ruban.**

Considérons l'exemple très simple suivant. On veut ajouter un nombre à un autre nombre ; on dispose des opérations suivantes :

— chargement de l'accumulateur	symbolisé par :		CHA
— addition au contenu de l'accu.	»	»	ADD
— rangement du contenu de l'accu.	»	»	RCA
— arrêt du programme	»	»	HLT

Il faut également disposer d'un moyen de réserver des mots mémoire pour les nombres à additionner ; supposons qu'on utilise le symbole DC suivi d'une constante. Le programme peut alors s'écrire ainsi :

CHA X	chargement du contenu de X dans l'accu.
ADD Y	addition du contenu de Y
RCA Y	rangement du résultat dans Y
HLT	arrêt
X/DC 25	définition des adresses X et Y et de leur
Y/DC 7	contenu initial

Nous voyons les symboles X et Y apparaître plusieurs fois de façon différente : — dans les instructions elles-mêmes, comme partie adresse de l'instruction et — comme définition d'adresse précédant le symbole DC. Chaque symbole peut apparaître autant de fois que cela est nécessaire comme partie adresse d'instruction mais doit apparaître une fois et une seule, comme définition d'adresse. Ainsi, si la première instruction du programme occupe l'adresse 0, et si chaque instruction occupe un mot, X représente l'adresse 4, Y l'adresse 5. Après l'exécution de ce programme, le contenu de X n'aura pas changé ; celui de Y sera remplacé par la somme des deux nombres, soit 32.

Supposons qu'on veuille réécrire ce programme de telle façon que les contenus des adresses X et Y ne soient pas modifiés en cours d'exécution. Il faut alors définir une autre adresse où l'on rangera le résultat. Le programme s'écrit alors ainsi :

```
 CHA X
 ADD Y
 RCA Z
 HLT
X/DC  25
Y/DC  7
Z/DC  0
```

Après exécution, le résultat se trouvera à l'adresse Z. Les contenus de X et Y auraient aussi pu être des résultats de calculs préalables ; dans ce cas la valeur numérique définie dans le programme est sans importance, mais il faudrait que la partie du programme calculant ces valeurs les range aux adresses correspondantes, au moyen d'instructions RCA X et RCA Y. Il faut remarquer la présence de l'instruction HLT à la fin du calcul ; en effet, sans cette instruction, le calculateur continuerait à prendre pour des instructions le contenu des adresses suivantes, ce qui a des conséquences imprévisibles mais souvent désastreuses dans un programme (destruction du programme en mémoire, par exemple).

Nous disposons ainsi d'une méthode pour écrire des programmes, d'une façon plus claire et plus pratique qu'en langage machine binaire, grâce à la possibilité d'utiliser des symboles. Cependant le programme doit être en binaire, au moment de le faire exécuter par le calculateur.

Traduction en langage machine

La traduction du programme symbolique en programme binaire est effectuée par un programme généralement appelé assembleur. Le travail de ce programme consiste essentiellement en deux points :

— Répertorier tous les symboles définissant des adresses dans le programme. Pour cela, l'assembleur lit une à une les instructions, en les comptant (nous supposerons pour simplifier qu'une instruction occupe toujours un mot mémoire). Chaque fois qu'il rencontre un symbole définissant une adresse il range ce symbole dans une table avec la valeur correspondante du comptage c'est-à-dire l'adresse.

— Remplacer, dans chaque instruction, le code opération symbolique par le code binaire correspondant. Les codes opérations sont obtenus à partir d'une table permanente, faisant partie de l'assembleur. Les adresses sont obtenues à partir de la table qui a été préalablement constituée. Les codes ainsi obtenus (opération et adresse) sont assemblés de façon à constituer une instruction machine.

L'assembleur, lorsqu'il effectue la traduction, occupe une partie de la mémoire du calculateur. D'autre part, les adresses ne sont pas toujours affectées définitivement à l'assemblage. En effet, s'il est toujours possible de définir une adresse à l'intérieur d'un programme, en convenant d'appeler adresse Ø l'adresse du premier mot occupé par le programme, on ne peut pas toujours choisir *a priori* l'adresse effective en mémoire à partir de laquelle on chargera le programme, pour le faire exécuter. Pour certains types de tâches, on utilise plusieurs programmes qui doivent résider simultanément en mémoire ; selon les besoins, ces programmes n'occuperont pas toujours les mêmes parties de la mémoire. De plus, l'assemblage d'un programme, surtout si ce programme est important, est une opération qui demande de la place en mémoire, et du temps de traitement. Il est alors plus rentable et plus commode de ne faire l'assemblage qu'une seule fois, et de disposer ensuite d'un programme binaire. Le programme binaire assemblé n'est donc pas directement chargé en mémoire centrale, mais est stocké par exemple sur bande ou disque magnétique, ou sur ruban perforé.

Chargement

Un programme ayant été traduit en binaire doit être chargé en mémoire centrale, pour être exécuté. On utilise pour cela un autre programme, appelé généralement *chargeur*. La nature des opérations effectuées par le chargeur dépend de deux choses.

— *La façon dont se fait l'adressage dans l'unité centrale, au cours de l'exécution.* Si l'adressage est fait par adresses, chaque instruction doit contenir l'adresse effective de l'opérande, qui peut dépendre de l'adresse initiale de chargement. Si l'adressage se fait par base et déplacement, toutes les adresses sont relatives, par rapport à une adresse de base définie dans le programme, et qui sera chargée dans le registre de base au début de l'exécution. Ces adresses relatives sont indépendantes du point de chargement.

— *Le type de code binaire généré par l'assembleur, dans le cas de l'adressage par adresses.* Si l'adresse de chargement a été choisie à l'écriture du programme, l'assembleur génère directement les adresses effectives. On dit qu'il s'agit d'un assembleur donnant un code binaire absolu. Le programme ainsi assemblé devra obligatoirement être chargé à l'adresse prévue. Par contre, si l'adresse de chargement doit rester au choix de l'utilisateur, l'assembleur ne peut pas générer les adresses effectives.

Ainsi dans l'exemple précédent, X représente l'adresse 4 dans le programme, mais si on choisit de le charger à partir de l'adresse 15ØØ, X devra représenter l'adresse 15Ø4 en mémoire. Selon les cas, le chargeur recopiera simplement en mémoire le code binaire, tel qu'il l'aura lu, ou bien calculera les adresses effectives, en ajoutant aux adresses définies dans le programme l'adresse de chargement, avant de mettre en place le code binaire. Dans ce dernier cas, le chargeur doit pouvoir faire la distinction, dans le code binaire généré par l'assembleur, entre les mots contenant des références mémoire, par exemple des instructions avec opérandes ou des pointeurs contenant des adresses, et qui doivent donc être modifiés en fonction de l'adresse de chargement, et d'autre part les mots contenant des constantes ou des instructions ne faisant pas appel à la mémoire et qui sont donc indépendants de l'adresse de chargement. Pour ce faire, l'assembleur « marque » chaque mot du programme différemment, selon qu'il devra ou non être modifié au chargement. Le chargeur ne modifie alors que les mots concernés.

Au moment du chargement, le chargeur occupe lui-même une partie de la mémoire. Il est donc impossible qu'un programme soit chargé à cet endroit, cependant on peut généralement utiliser cette place, par exemple pour stocker des données qui seront lues par le programme, le chargeur étant alors « écrasé ».

On se trouve apparemment dans un cercle vicieux : pour écrire et charger un programme, il faut disposer d'autres programmes. Ces programmes, comment les écrire et les charger ? Remarquons d'abord qu'il suffit d'obtenir au moins une fois ces deux programmes en binaire. On peut ensuite en faire autant de copies que l'on désire. Pour cette raison, ces programmes sont géné-

ralement réalisés par le constructeur de la machine et vendus en même temps. Le problème se réduit donc, pour le constructeur, à faire un assembleur et un chargeur, lorsqu'il fabrique un nouveau modèle. En général, on utilise pour cela un autre calculateur, avec un programme qui « simule » un assembleur, et qui produit les programmes binaires, par exemple sur ruban perforé.

Utilisation d'un calculateur

Le problème pratique qui se pose à l'utilisateur est la mise en œuvre du calculateur. Celle-ci dépend de la configuration physique de l'ensemble, en particulier des périphériques disponibles et de la capacité mémoire. Pour exécuter un programme, les opérations successives sont :

— Rédaction du programme et préparation du programme source (ruban perforé par exemple).

— Chargement et exécution de l'assembleur pour traduire le programme en binaire.

— Chargement et exécution du programme proprement dit.

Les opérations concernant directement le calculateur sont donc essentiellement le chargement et le démarrage des programmes. Elles peuvent être réalisées manuellement, en introduisant en mémoire, à l'aide des clés de la console un programme de chargement. En général, il s'agit d'un programme très court, destiné à lire un chargeur plus performant. Le démarrage s'effectue en chargeant de la même façon l'adresse de départ du programme. Ces opérations doivent être répétées pour chaque programme, il y a donc intérêt à les rendre automatiques.

Notion de système d'exploitation

L'adjonction au calculateur d'une mémoire de masse périphérique (par exemple une unité de disque magnétique) et l'utilisation de certains programmes *ad hoc* permettent de simplifier la tâche de l'utilisateur.

Pour cela, on réserve une partie de la mémoire centrale pour un programme appelé *moniteur* ou *superviseur;* ce programme est destiné à gérer l'ensemble des tâches en faisant tout ce qui ne nécessite pas une intervention particulière de l'utilisateur. Une partie du disque est également réservée au moniteur pour ses besoins propres, une autre partie comporte les programmes évoqués ci-dessus; le reste de la place est à la disposition des utilisateurs.

Le moniteur comporte plusieurs parties, assurant chacune certaines fonctions :

— communications avec les utilisateurs, par exemple par l'intermédiaire d'un télétype;

— décodage des informations fournies par l'opérateur;

— échanges entre unité centrale et périphériques.

Les informations enregistrées sur le disque sont groupées en fichiers. Un fichier est un ensemble structuré d'enregistrements, repéré par un *nom* grâce auquel le moniteur peut le localiser et l'utiliser. Certaines parties du moniteur sont résidentes, c'est-à-dire toujours présentes en mémoire centrale; les autres parties sont chargées uniquement en cas de besoin, et les parties de la mémoire utilisées sont de nouveau disponibles après. Ce procédé permet de ne réserver qu'une petite partie de la mémoire centrale au moniteur, le reste étant à la disposition des utilisateurs.

L'utilisateur dispose alors de commandes ou instructions de contrôle pour indiquer au moniteur les tâches à exécuter. Ces commandes sont passées soit par le clavier, soit encore par des cartes perforées, comportant un code spécial les identifiant comme cartes de contrôle.

Le démarrage proprement dit du calculateur consiste à charger manuellement un programme court (appelé *boostrap*) destiné à lire un programme d'initialisation. Ce programme charge la partie résidente du moniteur et lui passe le contrôle. Le moniteur est alors prêt à recevoir des commandes. Après cela il n'y a, en principe, plus d'opération à effectuer manuellement à la console, tout se déroulant sous le contrôle du moniteur.

A titre d'exemple, voici, de façon pratique, comment se déroule un travail sur un gros calculateur.

— On écrit le programme. Très souvent on utilise un langage de programmation plus évolué que le langage d'assemblage (voir plus loin).

— On perfore le programme sur cartes.

— On ajoute les cartes de contrôle donnant au moniteur toutes les indications concernant le travail (langage utilisé, exécution ou seulement traduction en langage machine, etc.).

— On place le tout dans le lecteur de cartes.

Le reste des opérations jusqu'à la sortie des résultats s'effectue sans autre intervention de l'utilisateur. Le moniteur et les autres programmes utilisés pour la réalisation d'un programme d'utilisateur constituent le *système d'exploitation*.

L'utilisation d'un système d'exploitation est nécessaire, car l'automatisation de beaucoup d'opérations, d'une part, réduit considérablement les risques d'erreurs, d'autre part, accélère le rendement de la machine par la suppression des interventions humaines, très lentes en comparaison de la vitesse du calculateur. Par ailleurs, elle facilite le travail des utilisateurs en les déchargeant de nombreuses opérations répétitives et fastidieuses.

Les gros calculateurs sont généralement dotés de systèmes puissants, permettant de gérer en même temps les travaux de plusieurs utilisateurs. Pour ces machines, il importe que le système soit performant, car il utilise lui-même une partie des ressources de l'ensemble (place en mémoire, temps d'unité centrale) et le coût d'exploitation de ces calculateurs est très élevé. On peut définir des critères de qualité, par exemple :

Rt = Rendement en temps d'unité centrale
Rm = Rendement en ressources mémoire

$$Rt = \frac{\text{Temps d'exécution des travaux d'utilisateur}}{\text{Temps total de travail}}$$

$$Rm = \frac{\text{Place mémoire disponible pour l'utilisateur}}{\text{Place mémoire totale disponible}}$$

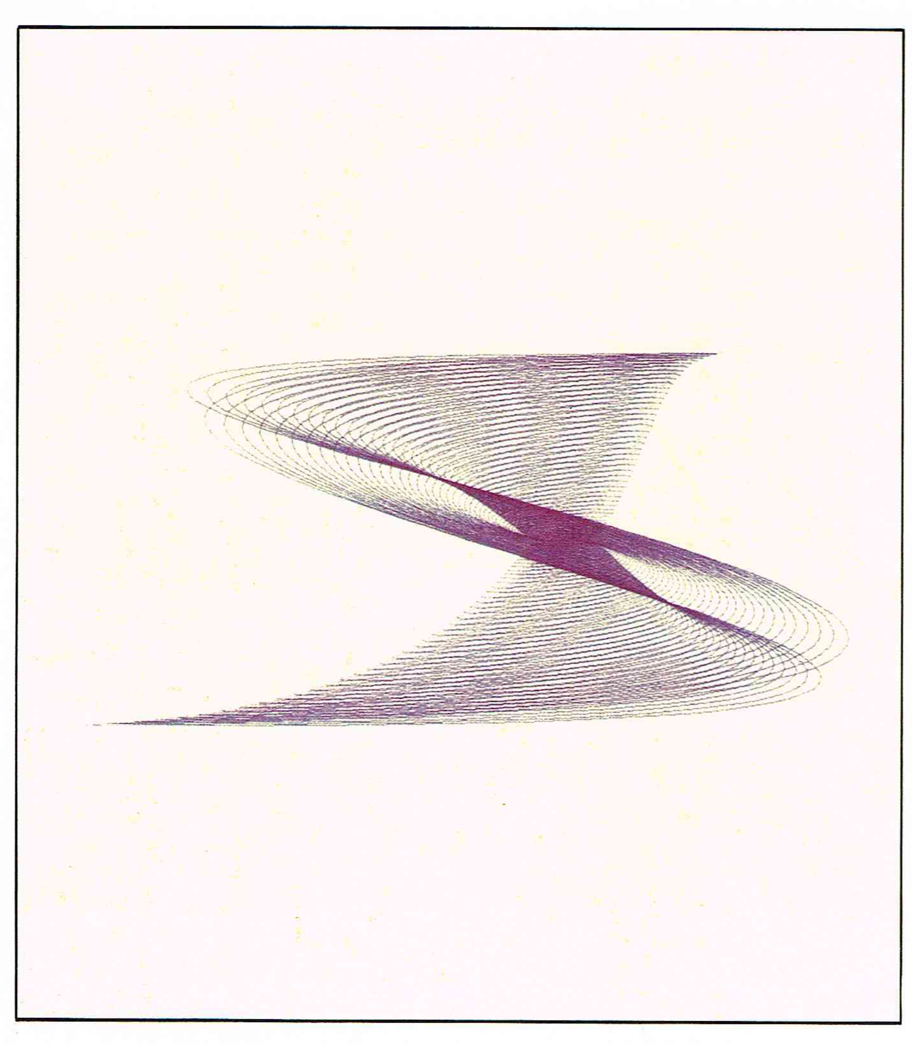
Un bon système d'exploitation doit offrir beaucoup de facilités aux utilisateurs, tout en ayant le meilleur rendement possible.

Il faut remarquer un point important : pour le calculateur, les programmes du système d'exploitation sont des programmes comme les autres. Le fonctionnement physique du calculateur ne dépend pas du programme en cours d'exécution. Par contre, pour l'utilisateur, ces programmes sont « transparents » car il les utilise en permanence, sans avoir à les connaître en détail. Le calculateur et le système

```
/********* PROGRAMME "MAX"                        PAL8-V9B  NO/DA/TE  PAGE 1

               /********* PROGRAMME "MAX"
               /          ***************
               /
               /CE PROGRAMME CHERCHE L'ADRESSE ET LA VALEUR DU PLUS GRAND
               /NOMBRE CONTENU DANS UNE TABLE DE 64 MOTS MEMOIRE, SITUEE
               /A PARTIR DE L'ADRESSE: TAB . LES NOMBRES SONT DES ENTIERS
               /POSITIFS.
               /LA VALEUR DU MAXIMUM EST RANGEE A L'ADRESSE : MAX
               /L'ADRESSE DU MAXIMUM EST RANGEE A L'ADRESSE : ADMAX
               /EN FIN DE TRAITEMENT LE PROGRAMME S'ARRETE
               /
       0200    *200
00200  7300            CLA CLL
00201  1227            TAD INPTR
00202  3230            DCA PTR
00203  7001            IAC
00204  3231            DCA MAX
00205  3232            DCA ADMAX
00206  1233            TAD LONG
00207  3234            DCA CNTR
00210  7100    ENCORE, CLL
00211  1231            TAD MAX
00212  7041            CIA
00213  1630            TAD I PTR
00214  7630            SZL CLA
00215  5222            JMP UPDATE
00216  2230    INCR,   ISZ PTR
00217  2234            ISZ CNTR
00220  5210            JMP ENCORE
00221  7402            HLT
00222  1230    UPDATE, TAD PTR
00223  3232            DCA ADMAX
00224  1630            TAD I PTR
00225  3231            DCA MAX
00226  5216            JMP INCR
00227  0235    INPTR,  TAB
00230  0000    PTR,    0
00231  0000    MAX,    0
00232  0000    ADMAX,  0
00233  7700    LONG,   -100
00234  0000    CNTR,0
00235  0000    TAB,    ZBLOCK 100
               $
```

▲ ***Exemple de programme assemblé. Ce programme a été écrit pour un calculateur PDP 8/E de Digital Equipment, et assemblé avec l'assembleur PAL 8. La liste du programme est imprimée par l'assembleur. Elle comporte cinq colonnes : 1. les adresses (en octal); 2. le contenu des mots mémoire correspondants (en octal); 3. les définitions d'adresses symboliques (par exemple, MAX représente l'adresse 231); 4. le code opération symbolique ou le contenu du mot; 5. l'adresse de l'opérande pour les instructions contenant une référence à une adresse. Tout ce qui est précédé du signe / constitue des commentaires et n'est pas pris en compte au cours de l'assemblage.***



▲ *Le graphisme par ordinateur offre des possibilités multiples : ici, tracé obtenu par intégration numérique d'un système d'équations différentielles.*

d'exploitation constituent un ensemble de ressources que l'utilisateur emploie pour ses besoins propres, à savoir l'exécution de ses programmes.

Les langages évolués

Nous avons vu que la programmation en langage machine, même symbolique, était affaire de spécialiste, car elle demande une bonne connaissance de la structure et du fonctionnement détaillé de la machine. De plus, un programme écrit pour un modèle de calculateur est inutilisable sur un autre modèle. Pour ces raisons, on a créé des langages de programmation plus évolués, ayant un formalisme plus en rapport avec les problèmes traités, et surtout indépendants dans leur syntaxe du type de machine utilisée, mettant ainsi la programmation à la portée de beaucoup plus de gens et d'une façon relativement facile.

Le premier langage évolué a été créé en 1956. Il s'agit de **FORTRAN** (FORmula TRANslator). Ce langage utilise des notations très semblables aux notations mathématiques habituelles. Il est très utilisé dans les disciplines scientifiques, car il est surtout orienté vers les calculs.

ALGOL : plus évolué que Fortran, donc plus performant, il est cependant peu utilisé avec les calculateurs. Il est utilisé comme moyen de communication entre mathématiciens pour décrire des algorithmes.

COBOL : langage orienté vers les travaux de gestion, il n'est pas destiné à faire des calculs très compliqués, mais est très développé du point de vue des entrées-sorties et de la gestion de fichiers de données.

PL1 : langage universel, permettant aussi bien les calculs importants que la gestion. Créé en 1970, il est encore moins utilisé que Fortran ou Cobol.

Un langage évolué utilise des symboles et des opérateurs. Ces opérateurs représentent des opérations abstraites, et non les opérations qu'il faut faire exécuter concrètement au calculateur. Ainsi, par exemple, l'instruction WRITE en Fortran est une instruction de sortie, mais demande en réalité des centaines d'instructions machine.

Pour chaque langage et chaque type de calculateur, il faut un procédé de traduction en langage machine. La traduction est faite par un programme spécifique au calculateur utilisé ; ce programme est appelé compilateur, son travail comporte deux parties.

— Pour chaque instruction du langage considéré, décomposer l'opération en une suite d'opérations élémentaires exécutables par la machine. C'est une tâche comparable à ce que fait un programmeur qui prépare un programme en langage machine symbolique.

— Générer le code binaire correspondant aux séquences ainsi créées. Il s'agit en fait d'assembler le programme. Il se trouve souvent que l'assembleur utilisé sur un calculateur soit une partie d'un compilateur.

Chaque langage possède des règles d'écritures pour que chaque instruction puisse être analysée correctement et conformément au désir du programmeur. Le résultat de cette analyse par le compilateur doit d'une part représenter l'instruction initiale, d'autre part correspondre à la réalité physique du calculateur. En d'autres termes, l'analyse doit aboutir à une description finale du programme uniquement à l'aide d'éléments du langage machine.

Emploi des ordinateurs

Comme moyen de calcul

L'emploi des ordinateurs comme moyen de calcul permet d'aborder et de résoudre des problèmes, en particulier dans les disciplines scientifiques, qui seraient quasiment impossibles à traiter autrement. En effet certains problèmes n'ont pas de solution analytique connue et ne peuvent être traités que par des méthodes d'approximation numérique. D'autre part, la solution analytique de certains problèmes, connue sur le plan théorique, demanderait le travail de plusieurs générations de mathématiciens à cause de la complexité des calculs, de leur longueur ou de la masse de données à traiter.

Gestion

En général, les problèmes de gestion ne demandent pas des calculs d'une grande complexité mathématique. Cependant, ils portent sur de grandes quantités de données. Les ordinateurs permettent de tenir à jour et d'exploiter efficacement et rapidement des masses énormes de données (gestion de stocks, comptabilité, banques de données, etc.).

Temps réel

Dans les applications précédentes, le calculateur est employé de façon isolée. Il peut aussi être intégré à un ensemble plus vaste. Il s'agit des applications dites « en temps réel ». Le calculateur est connecté à des organes de mesure ou de surveillance pour acquérir des données, les trier, les stocker, en traiter une partie en temps réel, c'est-à-dire au rythme des phénomènes physiques étudiés. Il peut, en fonction des résultats trouvés, agir sur des organes de commande. Un exemple d'application est la conduite automatique de processus industriel, où le calculateur pilote et surveille une installation de production sans intervention humaine directe, sauf situation exceptionnelle (panne, entretien, modifications de la production).

BIBLIOGRAPHIE

CORGE C., *Éléments d'informatique. Informatique et démarche de l'esprit,* Larousse. - CROCUS, *Systèmes d'exploitation des ordinateurs. Principes de conception,* Dunod. - DEBRAINE P., *Machines de traitement de l'information. Circuits et programmes,* Masson. - DONDOUX J., MARANO P.-H., MERLIN J.-C., *Introduction à l'informatique. Structure et programmation des ordinateurs,* Armand Colin. - DREYFUS, *Fortran IV,* Dunod. - LABORDE J., *Cours pratique de langage Algol,* Dunod. - LAURENT A., *Principes de programmation des ordinateurs,* Masson. - MEINADIER J.-P., *Structure et fonctionnement des ordinateurs,* Larousse. - DE PALMA R., *Cours moderne de calcul automatique,* Albin Michel. - DU ROSCOÄT J., *Conception de la programmation des ordinateurs,* Masson. - SIMON J.-C., *Introduction au fonctionnement des ordinateurs,* Masson.



Doisneau - Rapho

▲ *Le marché monétaire est le marché de l'argent au jour le jour. Les investisseurs institutionnels (grandes banques, compagnies d'assurances, S.N.C.F., etc...) y interviennent pour équilibrer leur trésorerie. La Banque de France y joue un rôle prépondérant par des transactions journalières importantes qui assurent l'équilibre de la balance des paiements du pays et le soutien de la monnaie nationale.*

MATHÉMATIQUES FINANCIÈRES

Le système monétaire est né le jour où les hommes ont privilégié un bien particulier (l'or ou n'importe quel métal précieux) dont les propriétés (rareté, sécabilité, brillance, etc.) étaient telles qu'il constituait un instrument universellement admis pour mesurer la valeur des autres biens et régler alors les échanges commerciaux. La monnaie était ainsi un bien semblable aux autres biens mais dont la *valeur d'usage* était augmentée d'une *valeur d'échange* reflétant la facilité offerte à son détenteur d'acquérir d'autres biens. Avec le développement des échanges commerciaux à partir du Moyen Age, on a vu cette valeur d'échange prendre une importance croissante au détriment de la valeur d'usage. Cette évolution s'est très nettement accentuée depuis la révolution industrielle du XIXe siècle qui a suscité l'organisation du système du crédit et l'apparition de structures bancaires et financières très puissantes. La monnaie est dès lors devenue l'objet de transactions sans que celles-ci soient liées à la contrepartie en or qui la définissait originellement : elles sont liées désormais à sa valeur d'échange, c'est-à-dire aux possibilités de consommation et d'investissement qu'elle permet.

Ces transactions s'établissent entre des prêteurs (banques, compagnies d'assurances, État, particuliers) et des emprunteurs (entreprises, banques, particuliers, etc.). Le prêteur est rémunéré selon des modalités admises par tous les intervenants. Ces modalités sont fixées par des règles de calcul très précises dont nous verrons qu'elles reflètent exactement la loi de l'offre et de la demande sur le marché des capitaux.

En rétribution du service rendu, l'emprunteur verse donc au prêteur une certaine somme, appelée *intérêt* et qui n'est autre que le loyer de l'argent prêté. L'intérêt est bien sûr proportionnel au montant du prêt. Mais son calcul est différent suivant sa durée. A cet égard nous distinguerons les opérations financières à court terme *(intérêt simple)* des opérations à long terme *(intérêt composé)*.

Opérations financières à court terme - Intérêt simple

On appelle opérations à court terme les prêts et les emprunts dont la durée n'excède pas l'année.

Le **taux d'intérêt** est l'intérêt qui rémunère le prêt de 1 F pendant un an ; il s'agit donc d'un taux annuel que l'on notera i. Il est le plus souvent représenté par un pourcentage (10,33 %, 13 %, etc.).

Considérons un prêt caractérisé par les éléments suivants :

— *montant du prêt :* C,
— *durée du prêt :* n jours ($n \leqslant 360$),
— *taux annuel :* i.

L'intérêt I que l'emprunteur devra verser au prêteur au terme de la durée du prêt est proportionnel à ces trois grandeurs. Il est donné par la formule :

$$\text{I} = C\,\frac{i \cdot n}{360} \qquad \textbf{(1)}$$

A l'échéance, l'emprunteur devra rembourser la somme de $C\left(1 + \frac{i \cdot n}{360}\right)$.

Exemple : C = 250 000 F $i = 12$ % $n = 90$ jours

$$\text{I} = \frac{250\,000 \times 12 \times 90}{360} = 7\,500 \text{ F}$$

La formule **(1)** exprime bien la proportionnalité de I au nombre de jours de la durée du prêt. Ceci est remarquable, compte tenu des opérations financières au jour le jour auxquelles se livrent les grands organismes économiques et financiers, notamment sur le marché monétaire. Ces opérations portent sur des montants considérables et le report d'un jour ou deux de l'échéance d'un prêt

▶ Page ci-contre, figure 2 : représentation schématique du principe d'équivalence. Figure 3 : l'évaluation d'une suite d'annuités dépend du taux d'évaluation, du montant de chaque annuité et du nombre de ces annuités.

▼▶ Ci-dessous, tableau I : valeur acquise d'un capital de 1 F au bout de n années. En bas, tableau II : valeur actuelle de 1 F disponible dans n années. Ci-contre, figure 1 : voir développement dans le texte.

donne lieu à des coûts financiers que le gestionnaire ou le banquier ne néglige pas. D'où l'importance des délais de paiement dans les opérations commerciales puisque tout retard ou toute avance dans les paiements a des conséquences sur la rentabilité de l'entreprise. C'est tout le problème de la gestion optimale de la trésorerie des firmes qui se trouve posé au travers de la formule **(1)**.

Notons que l'intérêt se paie à l'échéance du prêt, au moment du remboursement, et non avant. C'est ainsi que les opérations d'**escompte** sont pratiquées à l'encontre de cette règle puisque le banquier fait payer l'intérêt au début du prêt (en le déduisant du montant demandé). Il s'ensuit un coût financier plus élevé pour l'emprunteur. Autrement dit, le taux d'intérêt annoncé dans les opérations d'escompte (e) est inférieur au taux réellement pratiqué (t) :

Soit C le capital prêté à un terme de n jours. Le taux d'escompte annoncé par le banquier est e. Au moment du versement des fonds, celui-ci va verser au client non pas C mais $\left(C - \frac{ne}{360} C\right)$. Au bout des n jours, le client remboursera C. Le taux réellement pratiqué t est donné par la formule :

$$\left(C - \frac{ne}{360} C\right)\left(1 + \frac{nt}{360}\right) = C$$

soit $t = \dfrac{e}{1 - \dfrac{ne}{360}}$; on a donc $e < t$.

Opérations à long terme - Intérêts composés

Considérons un prêt durant plusieurs années (5, 10, 20 ans par exemple). Il est légitime d'admettre que l'intérêt produit au titre d'une année est ajouté au capital emprunté pour former un nouveau capital portant intérêt pendant les années suivantes. Tel est le principe de la règle des intérêts composés.

Valeur acquise par un capital au bout de n années

Soit D le montant du prêt et i le taux (annuel) auquel est rémunéré le prêt. Le prêteur cherche à connaître le montant A_n qui lui reviendra lorsque le prêt lui sera remboursé, en une seule fois, à l'échéance. A_n est appelée la *valeur acquise du capital D*.

A la fin de la première année, le prêt a produit un intérêt s'élevant à $i \cdot D$; la valeur acquise A_1 est égale à $D\,(1 + i)$. L'intérêt est alors ajouté au capital ; il devient alors productif d'intérêt pour l'année suivante. A la fin de la deuxième année, le prêt a produit un intérêt s'élevant à $i \cdot A_1$. C'est-à-dire que la valeur acquise au bout de deux ans est égale à $A_1\,(1 + i)$, soit $D\,(1 + i)^2$. En raisonnant de la sorte pour les années suivantes, on montre que la valeur acquise A_n est donnée par :

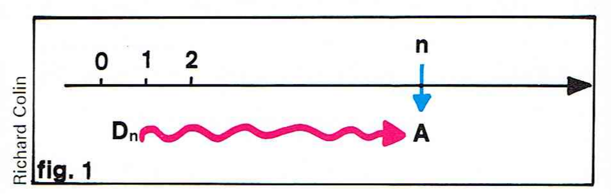
(2) $$A_n = A_{n-1}\,(1 + i) = D\,(1 + i)^n$$

Il ressort de la formule **(2)** que A_n croît « très vite » avec le nombre d'années n *(croissance* dite *« exponentielle »)*.

Exemple : 1 F placé en 1926 au taux de 7 % deviendrait aujourd'hui (en 1976) 29,45 F (tableau I). C'est-à-dire qu'un capital placé dans ces conditions aurait été multiplié par 30 ! (les effets des dévaluations successives de la monnaie depuis cette époque ont bien sûr largement corrigé cette progression...).

Valeur actuelle d'une somme disponible dans n années

La valeur actuelle d'une somme A payable dans n années est la somme D_n qui, placée à intérêts composés pendant ces n années au taux i, aurait pour valeur acquise A (*fig.* 1 et *tableau* II).



D'après ce que nous venons de voir au paragraphe précédent, nous avons :

$$D_n = \frac{A}{(1 + i)^n} = A\,(1 + i)^{-n}$$

Cette formule est d'une importance tout à fait essentielle dans le calcul économique et financier car elle permet de comparer et donc d'ajouter des sommes disponibles à des époques différentes.

Exemple : supposons une personne placée devant l'alternative suivante : recevoir 950 F dans 2 ans, ou recevoir 1 300 F dans 5 ans. Que va-t-elle choisir, sachant que le taux d'intérêt en vigueur sur le marché est de 10 % ? Pour répondre à la question, il faut calculer les valeurs actuelles de ces deux sommes : la valeur actuelle de 950 F est $\frac{950}{(1 + 0{,}10)^2}$ celle de 1 300 est de $\frac{1\,300}{(1 + 0{,}10)^5}$ soit respectivement 813,50 F et 807,17 F. Le meilleur choix est donc le premier.

On aurait tout aussi bien pu comparer les valeurs acquises pour ces deux sommes au moment de la 5e année, ce qui revient à comparer 950 $(1 + 0{,}10)^3$ et 1 300. Le résultat aurait été le même.

Tableau I - Valeur acquise d'un capital de 1 F au bout de n années en fonction du taux d'intérêt i

n \ i	9,25 %	9,50 %	9,75 %	10 %
1	1,092 500	1,095 000	1,097 500	1,100 000
2	1,193 556	1,199 025	1,204 506	1,210 000
3	1,303 960	1,312 932	1,321 946	1,464 100
4	1,424 577	1,437 661	1,592 292	1,610 510
10	2,422 225	2,478 228	2,535 393	2,593 742
50	83,381 991	93,477 253	104,767 463	117,390 850

Tableau II
Valeur actuelle de 1 F disponible dans n années

n \ i	9,25 %	9,50 %	9,75 %	10 %
1	0,915 332	0,913 242	0,911 162	0,909 091
2	0,837 832	0,834 011	0,830 216	0,826 446
3	0,766 895	0,761 654	0,756 461	0,751 315
4	0,701 963	0,695 574	0,689 258	0,683 013
10	0,412 844	0,403 514	0,394 416	0,385 543
50	0,011 993	0,010 698	0,009 545	0,008 519

Ainsi donc, d'une manière générale, pour comparer deux capitaux disponibles à des dates différentes, il suffit de comparer leurs valeurs (actuelles ou acquises) à une *date identique.*

Équivalence

Deux capitaux ou deux ensembles de capitaux sont équivalents à une date donnée et à un taux donné si, à cette date, leurs valeurs exprimées en fonction de ce taux sont égales. Il s'agira de valeurs acquises ou actuelles suivant que la date d'évaluation est antérieure ou postérieure aux dates auxquelles sont disponibles les deux capitaux.

Il peut en effet arriver que, pour des raisons commerciales ou financières, il soit nécessaire de remplacer un capital ou un ensemble de capitaux par un autre capital ou un autre ensemble de capitaux. Pour que l'échange ne soit contesté par aucune des parties en présence (débiteur et créancier), il faut qu'il se fasse entre des capitaux équivalents de manière à ce que le *temps* ne soit point négligé.

Exemple : un commerçant a contracté deux dettes auprès d'un même créancier : la première en 1972 d'un montant de 80 000 F à 8 % devant échoir en 1979, la seconde en 1974 d'un montant de 100 000 F à 9 % devant échoir en 1981. Ces deux dettes sont remboursables en une seule fois à échéance. Le commerçant se met d'accord aujourd'hui (1976) avec son créancier pour remplacer ces deux paiements par un paiement unique à échéance en 1980 au taux de 10 %. Quel doit être le montant de cette nouvelle dette (*fig.* 2) ?

En 1979, la première dette vaudra 80 000 $(1 + 0{,}08)^7$, la deuxième en 1981 vaudra 100 000 $(1 + 0{,}09)^7$. Les valeurs actuelles en 1976 de ces deux sommes, évaluées au taux de 0,10 %, sont respectivement de :

$$\frac{80\,000\,(1 + 0{,}08)^7}{(1 + 0{,}10)^3} \quad \text{et} \quad \frac{100\,000\,(1 + 0{,}09)^7}{(1 + 0{,}10)^5}.$$

Le montant X cherché est égal à la somme de ces deux nombres : X = 96 965 + 113 336 = 201 301 F.

Les annuités

On appelle *annuités* des versements dont les échéances sont séparées par des intervalles égaux (en général d'une année). Elles sont constamment utilisées en finance, soit dans le remboursement d'une dette, soit dans la constitution d'un capital ; il importe donc de pouvoir évaluer une suite d'annuités. Cette évaluation va dépendre du taux d'évaluation, du montant de chaque annuité et du nombre de ces annuités.

Considérons donc une séquence de n annuités d'un montant $a_1, a_2, ..., a_k, ..., a_n$. Par convention, on posera que l'annuité a_k échoit à la fin de l'année k (*fig.* 3).

L'évaluation d'une telle suite est une application du principe d'équivalence que nous avons présenté. A la date 0, la valeur V_0 de cette suite est égale à la somme des valeurs actuelles des annuités a_k (on parle de *somme actualisée*).

$$V_0 = a_1\,(1 + i)^{-1} + a_2\,(1 + i)^{-2} + ... + a_k\,(1 + i)^{-k} + ... + a_n\,(1 + i)^{-n}.$$

Le cas que l'on rencontre le plus fréquemment est celui où les annuités sont égales : $a_k = a$ $(k = 1, ..., n)$; V_0 s'écrit alors

$$V_0 = a\,[(1 + i)^{-1} + ... + (1 + i)^{-k} + ... + (1 + i)^{-n}]$$

Le second membre de cette expression est la somme d'une progression géométrique de premier terme $a\,(1 + i)^{-1}$ et de raison $(1 + i)^{-1}$. A l'aide d'une formule bien connue, on écrit :

$$\textbf{(3)} \qquad V_0 = a\,\frac{1 - (1 + i)^{-n}}{i} \qquad (\textit{tableau III})$$

Si l'on connaît V_0 et si l'on cherche la valeur de l'annuité a, on obtient :

$$\textbf{(4)} \qquad a = \frac{V_0 \cdot i}{1 - (1 + i)^{-n}} \qquad (\textit{tableau IV})$$

Il peut être également intéressant de connaître la valeur acquise V_n de la suite d'annuités au moment du dernier versement : elle est égale à la somme des valeurs acquises par les différentes annuités :

$$V_n = a_1\,(1 + i)^{n-1} + a_2\,(1 + i)^{n-2} + ... + a_n$$

soit : $V_n = V_0\,(1 + i)^n$.

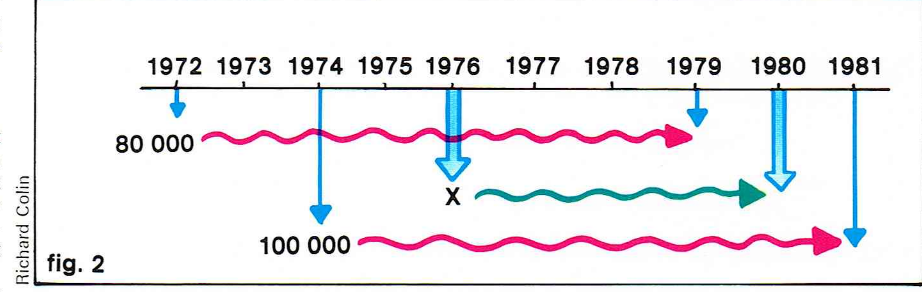


fig. 2

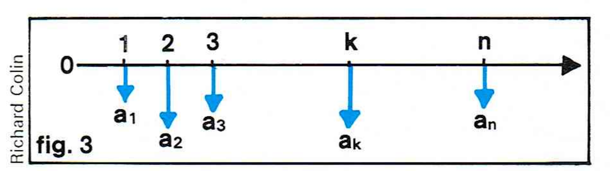


fig. 3

▼ ***Tableau III : valeur actuelle d'une suite d'annuités de 1 F versées pendant n années. Tableau IV : valeur des annuités constantes qui amortissent en n années un capital de 1 F.***

Tableau III - Valeur actuelle d'une suite d'annuités de 1 F versées pendant n années (cf. formule 3)

n \ i	9,25 %	9,50 %	9,75 %	10 %
1	0,915 332	0,913 242	0,911 162	0,909 091
2	1,753 164	1,747 253	1,741 377	1,735 537
3	2,520 059	2,508 907	2,497 838	2,486 852
4	3,222 022	3,204 481	3,187 096	3,169 866
10	6,347 637	6,278 798	6,211 116	6,144 567
50	10,681 157	10,413 708	10,158 513	9,914 815

Tableau IV - Valeur des annuités constantes a qui amortissent en n années un capital de 1 F (cf. formule 4)

n \ i	9,25 %	9,50 %	9,75 %	10 %
1	1,092 500	1,095 000	1,097 500	1,100 000
2	0,570 397	0,572 327	0,574 258	0,576 190
3	0,396 816	0,398 580	0,400 346	0,402 115
4	0,310 364	0,312 063	0,313 765	0,315 471
5				0,263 797
10	0,157 539	0,159 266	0,161 002	0,162 745
50	0,093 623	0,096 027	0,098 440	0,100 859

Autrement dit, la valeur acquise de la suite d'annuités est égale à la valeur acquise de sa valeur actuelle.

Lorsque les annuités sont constantes, on obtient la formule :

$$V_n = a \frac{(1+i)^n - 1}{i}$$

Application aux emprunts

Emprunt indivis

Un emprunt indivis est une opération financière qui met en lice *un* emprunteur et *un* créancier : le créancier prête un capital K à son débiteur, lequel lui rembourse ce capital augmenté des intérêts sous forme d'annuités (que nous supposerons constantes pour la commodité du calcul). Comment fixe-t-on le montant de l'annuité, connaissant le taux *i* auquel est rémunéré le prêt ? Celui-ci est calculé de telle façon que la *valeur actuelle de la suite des versements soit égale au montant du prêt :* grâce à la formule **(4)** on écrit :

$$\textbf{(5)} \qquad a = \frac{Ki}{1 - (1+i)^{-n}}$$

Ce mode de calcul est cohérent par rapport aux règles données plus haut. Cependant pour bien le comprendre, il est nécessaire de raisonner en *valeur acquise*.

Supposons que le créancier comme le débiteur aient tous deux la possibilité d'emprunter et de prêter de l'argent au taux *i*. Quels sont alors les comportements du créancier et du débiteur ?

— Le *créancier* a le choix entre deux possibilités :

* ne pas prêter le capital K à l'emprunteur en question et préférer le placer sur le marché pendant *n* années au taux *i* avec un remboursement en une seule fois au terme de ces *n* années ; il récolterait alors $K(1+i)^n$;

* accepter la proposition de l'emprunteur, ce qui lui rapporte chaque année une annuité *a*, soit au bout de *n* années $V_n = a(1+i)^{n-1} + a(1+i)^{n-2} + ... + a_n$, c'est-à-dire la valeur acquise de la suite d'annuités.

Il n'acceptera cette proposition que si $V_n \geqslant K(1+i)^n$.

— Le *débiteur* a lui aussi le choix entre deux possibilités :

* accepter la proposition du prêteur d'un prêt remboursable en *n* annuités ;

* préférer emprunter le capital K sur le marché et rembourser en une seule fois $K(1+i)^n$ au bout des *n* années ; mais il aurait alors la possibilité, sans que cela le gêne plus que dans la première possibilité, d'en consacrer chaque année une partie *a* à un placement au taux *i*, placement pouvant durer jusqu'à la *n*-ième année !

Il se livrera à cette double opération (emprunter d'une part pour replacer de l'autre) si $V_n \geqslant K(1+i)^n$. Par conséquent il n'acceptera la proposition du créancier que dans le cas où $V_n \leqslant K(1+i)^n$.

La loi de l'offre et de la demande conduit donc débiteur et créancier à se mettre d'accord sur l'égalité :

$$V_n = K(1+i)^n = V_0(1+i)^n \text{ soit } K = V_0$$

d'où la justification de la formule **(5)**.

L'annuité est ainsi déterminée par deux éléments : le remboursement du capital et le paiement des intérêts sur le capital non encore remboursé (l'*amortissement* et l'*intérêt*).

Dans la pratique, il importe de distinguer ces deux éléments. Du point de vue de l'entreprise, ils ne sont pas assujettis au même régime fiscal, les intérêts étant déductibles de la base imposable (ils entrent dans la catégorie des frais financiers) et le remboursement du capital étant pris en compte au niveau des amortissements.

Soit donc A_k l'amortissement inclus dans la *k*-ième annuité, D_k la somme restant due (en capital) immédiatement après le versement de la *k*-ième annuité. Le débiteur emprunte donc une somme $K = D_0$ qui lui est versée au début de la première année. A la fin de la première année, il paie à son créancier une première annuité a_1, celle-ci comprend l'intérêt produit par le placement du capital D_0 pendant une année au taux *i*, soit $D_0 i$; la différence $a_1 - iD_0$ représente la part du capital A_1 remboursée à la fin de la première année.

Le capital restant dû pendant la deuxième année est donc $D_1 = D_0 - A_1$; c'est donc sur D_1 que sera calculé l'intérêt compris dans la deuxième année. D'où $a_2 = iD_1 + A_2$ et ainsi de suite... jusqu'à la *n*-ième année.

On représente cette suite de calculs dans un tableau (*tableau* V).

A l'issue de la *n*-ième année, le capital est intégralement remboursé et donc $D_n = 0$. Ainsi donc $A_n (1+i)$.

Remarques :

— le taux d'intérêt est supposé constant pendant toute la durée de l'emprunt ;

— les intérêts dus *au titre de l'année écoulée* sont obligatoirement versés puisqu'ils sont compris dans l'annuité ; ils ne donnent donc pas lieu à capitalisation.

Grâce aux relations entre a_k, D_k et A_k dans le *tableau* V, il est possible d'écrire une relation entre deux annuités consécutives et les amortissements qui y figurent. Nous avons :

$$a_k = D_{k-1}\, i + A_k ; \qquad D_k = D_{k-1} - A_k ;$$
$$a_{k+1} = D_k\, i + A_{k+1} ;$$

d'où l'on tire $a_k - a_{k+1} = (D_{k-1} - D_k)\, i + A_k - A_{k+1}$

$$\textbf{(6)} \qquad a_k - a_{k+1} = A_k(1+i) - A_{k+1}$$

Emprunt à annuités constantes

Lorsque $a = a_k$, la formule **(6)** permet d'écrire : $A_{k+1} = A_k(1+i)$. Ainsi donc, la suite des amortissements est une progression géométrique de raison $(1+i)$ et dont le premier terme est $A_1 = a - iD_0$. Sachant que $a = \frac{D_0 i}{1-(1+i)^{-n}}$ il est facile de vérifier que la somme des amortissements $\sum_{i=1}^{n} Ai$ est égale au capital prêté D_0.

Ces quelques règles suffisent pour établir le *tableau d'amortissement d'un prêt* (*tableau* VI) : il s'agit d'un tableau qui, année par année, indique l'intérêt payé, la part de capital remboursé ainsi que la somme restant due (dette non encore éteinte).

Soit, par exemple, un emprunt de 200 000 F sur 5 ans à 10 %. A l'aide du *tableau* VI, on détermine le montant de l'annuité : $a = 200\,000 \times 0{,}263\,797 = 52\,759{,}40$ F.

▼ *Ci-dessous, tableau V : calculs permettant de fixer le montant des annuités jusqu'au remboursement intégral du capital. En bas, tableau VI : tableau d'amortissement d'un prêt de 200 000 F sur 5 ans à 10 %.*

Tableau V

0		
1	$a_1 = D_0 i + A_1$	$D_1 = D_0 - A_1$
2	$a_2 = D_1 i + A_2$	$D_2 = D_1 - A_2$
3		
k	$a_k = D_{k-1} i + A_k$	$D_k = D_{k-1} - A_k$
n	$a_n = D_{n-1} i + A_n$	$D_n = D_{n-1} - A_n = 0$

Tableau VI
Amortissement d'un prêt de 200 000 F sur 5 ans

Années	Somme restant due	Intérêt	Amortissement	Annuités a
1	200 000,00	20 000,00	32 759,40	52 759,40
2	167 240,60	16 724,00	96 035,40	52 759,40
3	131 205,20	13 120,52	39 638,88	52 759,40
4	91 566,32	9 156,63	43 602,77	52 759,40
5	47 963,55	4 796,85	47 963,55	52 759,40
			200 000,00	

Emprunt - Obligations

Une entreprise (ou une administration) est souvent amenée à emprunter une somme importante pour procéder à de nouveaux investissements. Pour cela il est fréquent qu'elle fasse appel non pas à un seul prêteur mais à plusieurs. L'emprunt est alors divisé en *obligations* (nominales ou au porteur). La détention de celles-ci donne droit au versement d'un intérêt annuel contre remise d'un coupon. Chaque année, un certain nombre d'obligations sont remboursées, en général par tirage au sort. Ce nombre est déterminé par le tableau d'amortissement (*tableau* VII).

Exemple : considérons une entreprise qui émet un emprunt obligataire de C = 200 000 000 F réparti en 200 000 obligations de 1 000 F, à un taux de 10 % sur 5 ans en annuités constantes. Au terme de la k-ième année, l'annuité comprend une part d'intérêt et une part de remboursement A_k du capital. Cet amortissement est réparti en obligations qui sont alors remboursées. Le nombre N_k de titres remboursés est égal à $\frac{A_k}{C}$.

Il arrive parfois que le prix d'émission (E) des obligations soit inférieur à leur valeur nominale (C) et que le prix de remboursement (R) lui soit supérieur. Dans ce cas, le taux réellement pratiqué est plus élevé que le taux annoncé. Le principe de calcul des N_k reste sensiblement le même.

Application à la mesure de la rentabilité des investissements

Dans l'entreprise, l'investissement recouvre un grand nombre d'activités et de préoccupations : acquérir un nouvel outillage, organiser la formation de telle ou telle catégorie du personnel, développer un réseau commercial, acquérir une participation dans le capital d'une autre entreprise sont des activités que l'on appelle des investissements. Malgré leur diversité, elles partagent en commun le fait qu'elles produisent leurs effets sur plusieurs années aussi bien dans les charges dont elles grèveront les finances de l'entreprise que dans les recettes qu'elles occasionneront.

L'appréciation d'un investissement suppose que l'on ait à sa disposition une procédure d'arbitrage entre le présent et le futur de manière à comparer des entrées ou des sorties d'argent à des dates différentes. A côté des méthodes classiques permettant de mesurer la rentabilité d'un investissement, qui procèdent de techniques comptables (calculs de *ratios*) et qui donc ne font guère intervenir le temps, il existe plusieurs méthodes fondées sur l'actualisation ; il s'agit principalement de la *méthode de la valeur actuelle nette (V.A.N.)* et de celle du *taux de rendement interne (T.R.I.)*.

On suppose que la firme qui étudie un projet d'investissement connaît avec certitude le nombre n d'années de durée de l'investissement, les dépenses D_k et les recettes R_k occasionnées par l'investissement lors de la k-ième année, $0 \leqslant k \leqslant n$.

On appelle *cash-flow* de la k-ième année la différence (algébrique) $C_k = R_k - D_k$; on convient, dans ce type d'analyse, de ne prendre en compte que les encaissements et décaissements effectivement réalisés (on ne prend en compte ni l'amortissement du matériel ni les provisions de toutes sortes) ; C_k peut être positif ou négatif suivant les années. Le plus souvent la suite des cash-flows est négative pour les premières années puis positive pour les suivantes (on parlera alors d'investissement conventionnel).

La méthode de la valeur actuelle nette

La valeur actuelle nette d'un investissement est la somme (algébrique) des valeurs actuelles de tous les cash-flows, calculées avec un taux i dit *taux d'actualisation,* dûment choisi :

$$\text{(7)} \qquad \text{V.A.N.} = \sum_{k=0}^{n} C_k (1 + i)^{-k}$$

Exemple : soit l'investissement caractérisé par la suite des cash-flows : $C_0 = -14\,000$; $C_1 = +4\,000$; $C_2 = +6\,000$; $C_3 = +8\,000$. Sa valeur actuelle nette à 10 % est (voir *tableau* II) :

$$\text{V.A.N.} = -14\,000 + 4\,000 \times 0{,}909\,0 + 6\,000 \times 0{,}826\,7 + 8\,000 \times 0{,}751\,3 = +595{,}2 \text{ F.}$$

Tableau VII - Amortissement d'un emprunt de 200 000 obligations sur 5 ans

Années	Somme restant due	Amortissement	Nombre de titres remboursés
1	200 000 000	32 759 400	32 760
2	167 240 600	36 035 400	36 035
3	131 205 200	39 638 880	39 639
4	91 566 320	43 602 770	43 602
5	47 963 050	47 963 550	47 964
			200 000 titres

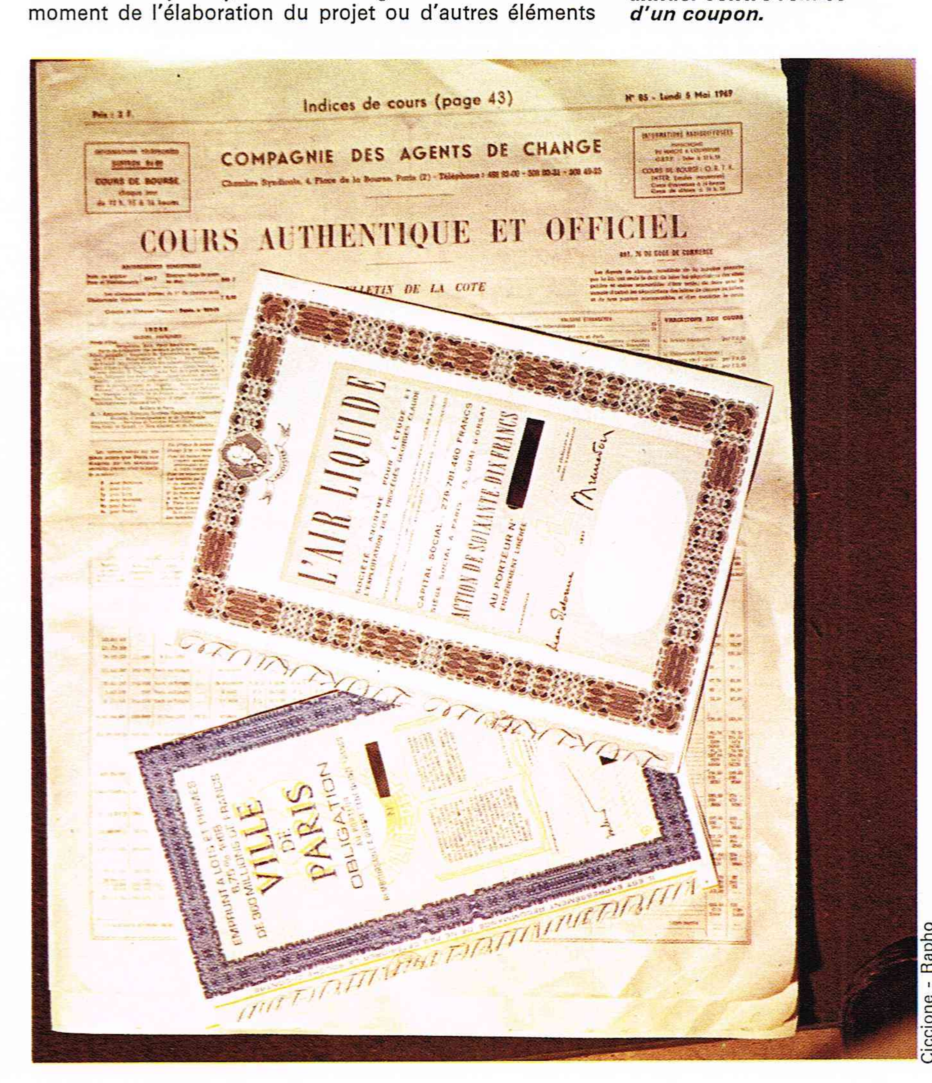
▲ *Tableau VII : tableau d'amortissement d'un emprunt obligataire donnant lieu, chaque année, au remboursement d'un certain nombre d'obligations.*

Dans cet exemple, on suppose que le coût de l'investissement (14 000 F) est intégralement supporté avant la première année.

La méthode de la V.A.N. consiste donc à classer les projets d'investissement dans l'ordre décroissant de leurs valeurs nettes. Le projet correspondant à la plus élevée sera réputé le meilleur.

Le choix du taux d'actualisation pose de multiples problèmes. Il peut dépendre de la structure financière de la firme (il pourra être égal à ce que les financiers appellent le « coût du capital »), des taux en vigueur sur le marché au moment de l'élaboration du projet ou d'autres éléments

▼ *L'obligation peut être nominale ou au porteur ; sa détention donne droit au versement d'un intérêt annuel contre remise d'un coupon.*



Ciccione - Rapho

plus spécifiques. Ce choix relève de l'analyse financière proprement dite.

Taux de rendement interne

La valeur nette d'un investissement est une fonction V.A.N. (i) du taux d'actualisation. La formule **(7)** indique même qu'il s'agit d'un polynôme de degré n en $(1+i)^{-1}$.

D'autre part V.A.N. $(0) = \sum_{k=0}^{n} C_k$ est positif (sinon l'investissement serait totalement improductif) ; lorsque $i \to +\infty$ V.A.N. (i) $\to C_0$ qui est négatif si l'on a affaire à un investissement conventionnel. Dans un tel cas, on montre que la fonction V.A.N. (i) est une fonction décroissante de i (pour $i \geqslant 0$). Prenant des valeurs positives et des valeurs négatives, il est clair qu'elle s'annule pour une certaine valeur (unique) i_0 du taux d'actualisation (*fig.* 4).

Le taux i_0 est appelé *taux de rendement interne de l'investissement*. Plus ce taux est élevé, plus l'on dira que l'investissement est rentable. On classera donc des projets par ordre croissant de taux de rendement interne.

La détermination de cette valeur i_0 pose quelques problèmes de résolution dans la mesure où V.A.N. (i) est un polynôme de degré n et que la recherche d'une racine ne peut se résoudre que par approximations successives (pour $n \geqslant 3$).

Comparaison des deux méthodes

La méthode du T.R.I. ne tient pas compte de la dimension de l'investissement : en effet si un investissement est caractérisé par une suite de cash-flows $\{C_k\}$, l'investissement ayant une suite de cash-flows $\{10\ C_k\}$ ou $\{1\ 000\ C_k\}$ aura le même taux de rendement interne. La méthode de la V.A.N., sur ce point, permet de prendre en compte la dimension de l'investissement.

Cependant la méthode du T.R.I. fournit un *critère objectif* de rentabilité alors que la méthode de la V.A.N. suppose la connaissance d'un taux d'actualisation dont la définition, ainsi que nous l'avons dit, n'est pas exempte d'arbitraire et de subjectivité.

Enfin la méthode de la V.A.N. permet de tenir compte de l'éventualité d'une variation de taux et donc de mesurer la rentabilité d'un investissement en fonction de diverses situations économiques possibles.

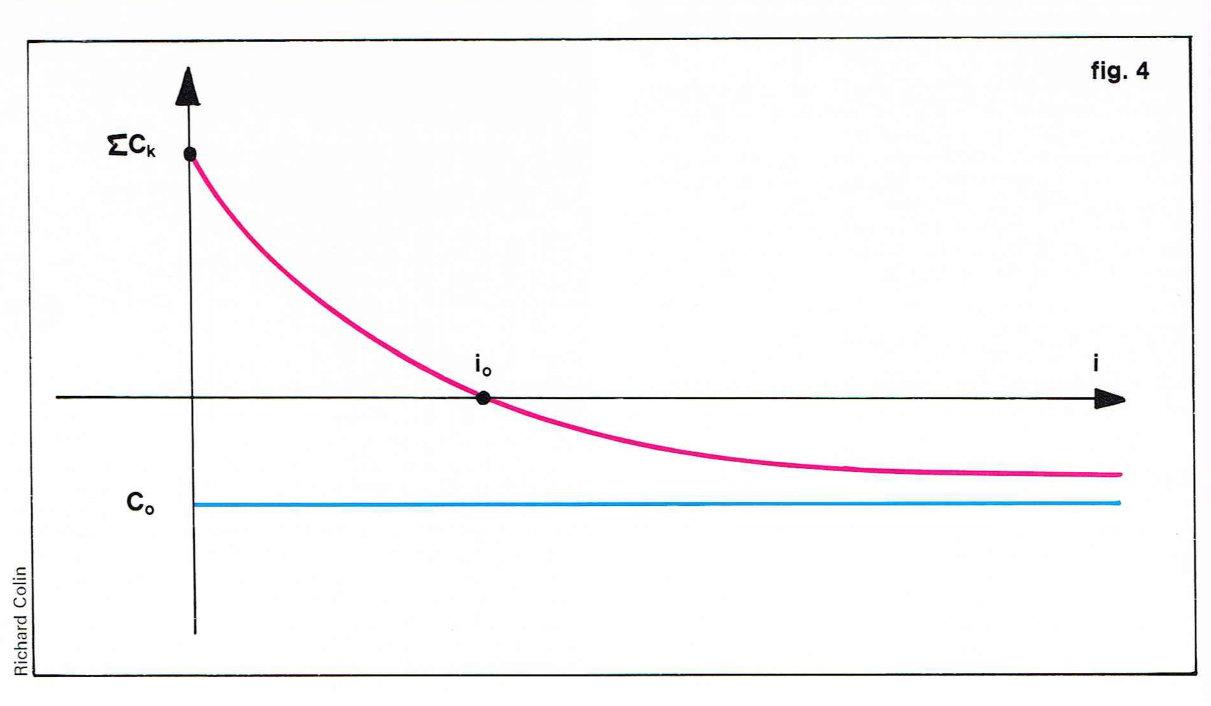
BIBLIOGRAPHIE

COLASSE B., *la Rentabilité, Analyse, Prévision et Contrôle,* coll. Entreprise, Dunod, 1973. - DUBRULLE L. et MAZERAN J., *Actualisation et Équivalence,* Masson, 1974. - PASCAL-FALGUIÈRES M., *Nouvelles Tables financières,* Foucher, 1970.



F. Hidalgo - TOP

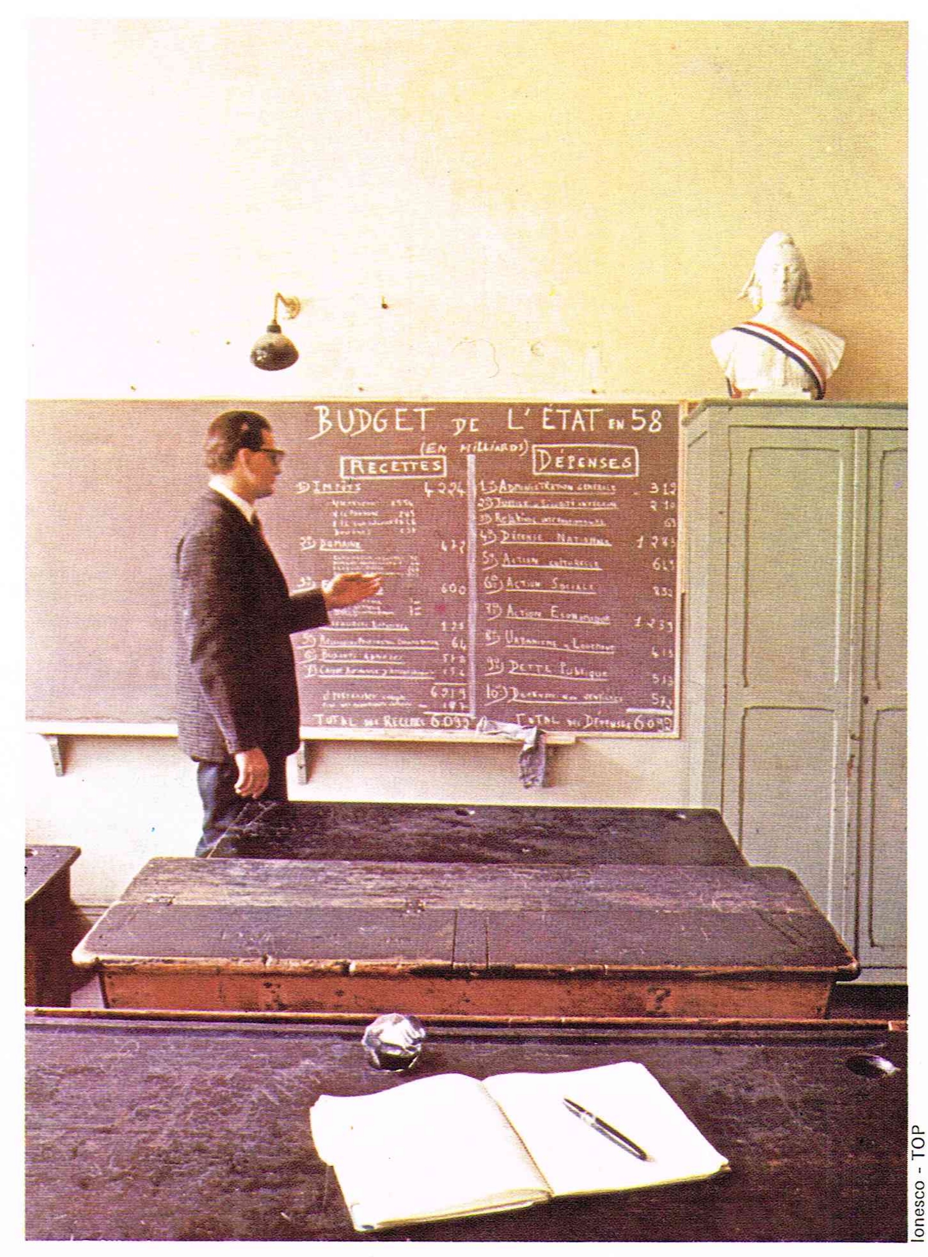
▲ *La monnaie est l'objet de transactions s'établissant entre des prêteurs (banques, État, particuliers) et des emprunteurs (entreprises, banques, particuliers), moyennant un intérêt proportionnel au montant du prêt et à sa durée.*



Richard Colin

▶ *Figure 4 : la valeur actuelle nette d'un investissement est une fonction V.A.N. (i) du taux d'actualisation.*

MATHÉMATIQUES ET SCIENCE ÉCONOMIQUE



▲ *Les mathématiques interviennent en économie comme démarche organisatrice de toute méthode scientifique, et pas seulement en tant qu'arsenal de recettes de calcul et de techniques quantitatives.*

La science économique est une science récente. A proprement parler, elle n'a guère que 150 ans ; cela est bien peu à côté de la physique d'Archimède, de la médecine d'Hippocrate... Cependant, les hommes n'ont pas attendu la deuxième moitié du XIX[e] siècle pour exercer leur intelligence à la compréhension de la réalité économique. L'économie politique est donc plus ancienne : littéraire et volontiers rhétorique, elle est avant tout de l'ordre du discours. Mais, au fil des siècles, elle s'est révélée peu apte à rendre compte de phénomènes économiques dont la complexité évoque celle des phénomènes physiques et par là même appelle une démarche scientifique et rigoureuse, c'est-à-dire l'élaboration de modèles.

Un **modèle** est une représentation abstraite et schématique du système économique que l'on veut analyser. Ce système comporte un grand nombre d'éléments sujets à des interactions, à des influences que l'économiste, *a priori*, connaît peu ou mal. Se fiant à son expérience ou à son intuition, il va privilégier les éléments du système qui lui semblent les plus essentiels à son fonctionnement. Ils sont alors *réduits à l'état d'objets mathématiques* (nombre, vecteur, ensemble, fonction, variable aléatoire, etc.). Quant aux liaisons de causalité et de dépendance qui existent entre eux et qui ont pu être mises en évidence, elles sont traduites en *relations mathématiques* (équations algébriques ou différentielles, inéquations, etc.).

Le modèle appelle alors une résolution mathématique (résolution de systèmes d'équations, problème d'optimisation, etc.). Le résultat est ensuite confronté avec le système. Au cours de cette réduction du réel, certains éléments ont été privilégiés, certaines hypothèses de comportement ont été posées. En outre, l'existence de solutions du modèle exige parfois que certaines hypothèses mathématiques soient introduites. Tout cela conditionne le résultat acquis. L'étude du modèle consiste donc également à voir dans quelle mesure ce résultat dépend de toutes ces hypothèses. Dès lors, il s'agit de mettre en évidence les *hypothèses critiques,* celles dont l'absence ou la modification perturbe le résultat.

On distingue deux types de modèles économiques : les modèles quantitatifs et les modèles qualitatifs.

Dans les *modèles quantitatifs* on s'attache à étudier les phénomènes mesurables, pour lesquels on dispose de *données,* recueillies, triées et interprétées à l'aide de méthodes statistiques (tests, sondages, analyse factorielle, etc.). Il s'agit de *modèles* dits *économétriques*.

Les *modèles qualitatifs* ont une tournure beaucoup plus théorique puisqu'ils traitent de phénomènes que l'on ne peut guère mesurer. Ces modèles sont plus justement appelés *modèles axiomatiques*. Bien qu'ils ne puissent pas faire l'objet de vérifications expérimentales, ils gardent une valeur descriptive et scientifique. La réalité (surtout en sciences humaines) n'est point seulement ce qui est mesuré ou mesurable.

Un modèle est donc, en tout état de cause, une représentation simplifiée, figée, voire caricaturale du système économique. Une contradiction apparaît ainsi inexorablement entre le modèle et la réalité qu'il prétend décrire. De l'approfondissement de cette contradiction émerge un nouveau modèle plus complet, prenant en compte un plus grand nombre d'éléments, établissant d'autres liaisons ou confortant les premières. L'étude de ce modèle fait apparaître à son tour une contradiction avec la réalité d'où émerge alors un nouveau modèle et ainsi de suite...

Les mathématiques sont ainsi d'une aide précieuse pour l'économiste puisque, par nature, elles le poussent à ne pas se satisfaire de son modèle, mais à définir de nouveaux objets, à établir de nouvelles liaisons plus générales. Elles suggèrent des voies de recherche qui dans de nombreux cas se sont révélées fécondes. Les mathématiques interviennent donc en économie avant tout en tant que démarche organisatrice de toute méthode scientifique et non point seulement comme un arsenal de recettes de calcul et de techniques dites quantitatives. Réduites à ce dernier rôle, les mathématiques restent d'un apport contestable en économie ; elles justifient dès lors un certain ésotérisme de bon aloi et participent d'un scientisme mystificateur. Par contre, restaurées en une démarche axiomatique, elles se révèlent être un instrument d'analyse et de synthèse étonnamment puissant parce que conscient de ses propres limites, c'est-à-dire apte, dans un mouvement dialectique, à les dépasser par une confrontation critique avec le réel.

Deux thèmes vont nous permettre d'illustrer le rôle joué par les mathématiques dans l'élaboration et dans la résolution de modèles axiomatiques : le problème de l'allocation optimale de ressources rares d'une part, et l'étude de la concurrence et des conflits de l'autre ; le premier sera éclairé par le formalisme de la programmation linéaire, le second par celui de la théorie des jeux.

Allocation optimale des ressources rares - Programmation linéaire

Position du problème - Exemple

L'entreprise Buchmoll et frères fabrique deux composés azotés : le chlorure d'ammonium (NH_4Cl) et le gaz ammoniac (NH_3) à partir de l'azote (N), de l'hydrogène (H) et du chlore (Cl). La fabrication d'une unité de chlorure d'ammonium nécessite donc une unité d'azote, quatre d'hydrogène et une de chlore, celle de gaz ammoniac, une d'azote et trois d'hydrogène.

L'entreprise dispose de 50 unités d'azote, 180 d'hydrogène et de 70 de chlore. Elle vend sur le marché le chlorure d'ammonium et le gaz ammoniac respectivement aux prix de 5 et 4 francs l'unité.

L'entreprise cherche à connaître le plan de production optimal, c'est-à-dire les quantités de NH_4Cl et de NH_3 qu'elle peut techniquement fabriquer compte tenu des ressources en matières premières (N, H, Cl) dont elle dispose et qui lui assurent le meilleur profit.

Appelons x_1 et x_2 les quantités produites de NH_4Cl et de NH_3. Cette production nécessite donc la transformation de $x_1 + x_2$ unités d'azote, de $4x_1 + 3x_2$ unités d'hydrogène et enfin de x_1 unités de chlore.

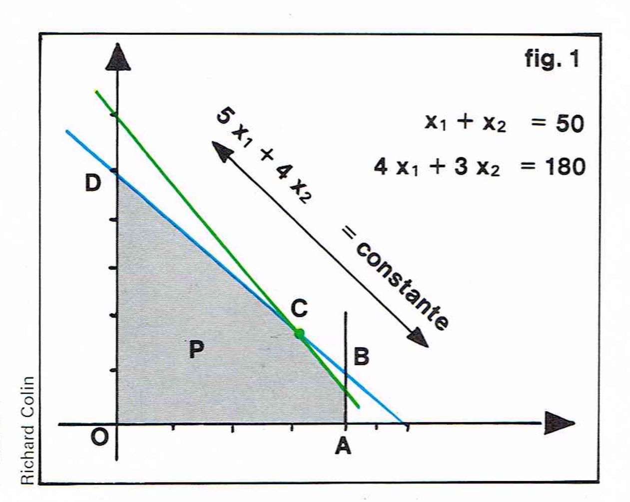
Compte tenu des disponibilités en ressources existantes, le problème que doit résoudre l'entreprise Buchmoll et frères se met sous la forme d'un problème d'optimisation :

$$\text{Max } z = 5x_1 + 4x_2$$

$$(1) \quad \text{sous les contraintes :} \quad \begin{cases} x_1 + x_2 \leqslant 50 \\ 4x_1 + 3x_2 \leqslant 180 \\ x_1 \leqslant 40 \\ x_1 \geqslant 0, \quad x_2 \geqslant 0 \end{cases}$$

Ce problème se résout graphiquement dans le plan $\mathbb{R}^2$ (*fig.* 1). L'ensemble des plans de productions techniquement réalisables $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ est représenté par le domaine P hachuré. Les droites d'équations $5x_1 + 4x_2 = \text{cte}$ sont parallèles. Le point C, intersection des droites d'équation $x_1 + x_2 = 50$ et $4x_1 + 3x_2 = 180$, est celui pour lequel la fonction $z = 5x_1 + 4x_2$ prend la plus grande valeur :

$$x_1^* = 30 \qquad x_2^* = 20 \qquad z^* = 230$$



▶ ***Figure 1 : résolution graphique du problème d'optimisation de l'entreprise Buchmoll et frères.***

Forme générale d'un programme linéaire

Le problème d'allocation optimale des ressources se présente de la façon suivante : l'entreprise dispose de p ressources rares (matières premières, équipements, main-d'œuvre, etc.) en quantités limitées ($b_1, ..., b_p$), d'où le qualificatif « rares ». A l'aide de celles-ci, elle peut fabriquer n produits différents dont les prix unitaires de vente sont fixés au préalable et égaux à $c_1, c_2, ..., c_n$. Dans ces conditions, quelles quantités $x_1, ..., x_n$ de ces n produits l'entreprise doit-elle fabriquer de manière à maximiser son gain ?

On suppose que pour fabriquer *une unité* du bien j, il faut a_{ij} unités de la ressource i ($a_{ij} = 0$ signifie que la ressource i n'entre pas dans la fabrication du bien j). On suppose en outre que la fabrication de x_j unités du bien j nécessite alors l'emploi de $x_j \cdot a_{ij}$ unités de la ressource i (hypothèse de rendements constants et de complémentarité stricte des facteurs de production). Enfin on suppose que l'entreprise ne produit pas de quantités négatives [$x_j \geqslant 0 \quad j = 1, ..., n$].

Un **plan de production** est donc un vecteur $\vec{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ de $\mathbb{R}^n$ que l'on dira *réalisable* et dès lors qu'il satisfera au système d'inéquations :

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n \leqslant b_1 \\ a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n \leqslant b_2 \\ \vdots \\ a_{p1}x_1 + a_{p2}x_2 + ... + a_{pn}x_n \leqslant b_p \\ x_1 \geqslant 0 \qquad x_2 \geqslant 0 \; ... \; x_n \geqslant 0 \end{cases}$$

Soit matriciellement, en posant

$$A = [a_{ij}]_{\substack{i=1,...,p \\ j=1,...,n}} \qquad \vec{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_p \end{bmatrix} :$$

$$\begin{bmatrix} A\vec{x} \leqslant \vec{b} \\ \vec{x} \geqslant \vec{0} \end{bmatrix}$$

le signe $\leqslant$ signifie en l'occurrence que l'inégalité a lieu pour toutes les composantes.

Parmi les plans de production réalisables, l'entreprise choisira celui qui maximise son profit représenté par la forme linéaire :

$$\langle c, \vec{x} \rangle = [c_1, ..., c_n] \cdot \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \sum_{j=1}^{n} c_j x_j$$

Un programme linéaire s'écrit sous la forme générale :

$$(2) \qquad \begin{array}{l} \text{Max } \langle c, \vec{x} \rangle \\ A\vec{x} \leqslant \vec{b} \\ \vec{x} \geqslant 0 \end{array}$$

La fonction $\mathbb{R}^n \to \mathbb{R} \; \vec{x} \to \langle c, \vec{x} \rangle$ s'appelle la *fonction objectif* du programme linéaire.

Réécrivons le programme **(2)** en introduisant p variables supplémentaires y_i, appelées *variables d'écart*. La variable y_i est associée à la i-ième contrainte de la manière suivante : l'inéquation $a_{i1}x_1 + ... + a_{in}x_n \leqslant b_i$ est équivalente au système :

$$\begin{cases} a_{i1}x_1 + ... + a_{in}x_n + y_i = b_i \\ y_i \geqslant 0 \end{cases}$$

On définit ainsi un vecteur $\vec{y}$ de $\mathbb{R}^n$ et le programme **(2)** s'écrit :

$$(3) \qquad \begin{array}{l} \text{Max } \langle c, \vec{x} \rangle + \langle \mathbf{O}, \vec{y} \rangle \\ A\vec{x} + \vec{y} = \vec{b} \\ \vec{x} \geqslant 0 \qquad \vec{y} \geqslant 0 \end{array}$$

avec $\mathbf{O} = (0, ... 0)$

Le programme linéaire est dit alors mis « sous forme canonique ».

Résolution d'un programme linéaire - Principe de la méthode du simplexe

On démontre mathématiquement que la solution optimale $\vec{x}^*$ du programme **(3)** n'appartient jamais à « l'intérieur » du domaine réalisable mais toujours à sa frontière, c'est-à-dire au simplexe défini par les contraintes linéaires (c'est-à-dire dans l'exemple de la *figure* 1 : au pentagone OABCD). La méthode du simplexe est une méthode itérative d'exploration des sommets du simplexe : à partir d'un plan de production réalisable (par exemple $x_1 = x_2 = ... = x_n = 0$; $y_i = b_i \; i = 1, ..., p$ si le vecteur $\vec{b}$ a toutes ses composantes positives), on détermine un autre plan de production, pour lequel la fonction objectif prend une valeur plus grande. Le passage d'un plan à l'autre s'opère par des transformations matricielles (portant à la fois sur la matrice A et sur le covecteur c) fondées sur la *méthode du pivot* (voir *Calcul numérique*). La procédure d'exploration s'arrête lorsque l'on ne peut plus améliorer la fonction objectif.

Programme dual

Revenons à l'exemple de l'entreprise Buchmoll : supposons qu'une autre entreprise, l'entreprise Dugommier et fils, veuille acheter les stocks d'azote, d'hydrogène et de chlore détenus par Buchmoll. Elle cherche alors à évaluer de manière optimale les prix u_1, u_2, u_3 auxquels elle devra acquérir ces stocks. L'entreprise Buchmoll n'acceptera de se livrer à cette transaction que si elle n'est pas déficitaire : autrement dit, si elle ne peut pas gagner plus en produisant le chlorure d'ammonium et le gaz ammoniac et en les vendant aux prix de 5 et 4 F ; c'est-à-dire si la *valeur* des matières premières nécessaires pour fabriquer une unité de NH_4Cl est supérieure à 5 F :

$$u_1 + 4u_2 + u_3 \geqslant 5$$

de même, pour NH_3, si elle est supérieure à 4 F :

$$u_1 + 3u_2 \geqslant 4$$

L'entreprise Dugommier, dans ces conditions, cherche à minimiser le coût de la transaction, c'est-à-dire à résoudre le programme linéaire suivant :

$$\begin{aligned} \text{Min } w &= 50\,u_1 + 180\,u_2 + 40\,u_3 \\ u_1 + 4\,u_2 + u_3 &\geqslant 5 \\ u_1 + 3\,u_2 &\geqslant 4 \\ u_1 \geqslant 0 \quad & u_2 \geqslant 0 \end{aligned}$$

Ce programme est appelé *programme dual* du programme **(1)** ; les variables u_1, u_2, u_3 sont les *variables duales* (et, partant, x_1, x_2 les *variables primales*).

La résolution graphique d'un tel programme est plus délicate (dans $\mathbb{R}^3$!) ; néanmoins, on montre que, à l'optimum $u_3 = 0$, la solution optimale est alors :

$$u_1^* = 1 \qquad u_2^* = 1 \qquad u_3^* = 0 \qquad w^* = 230$$

On remarque alors que $w^* = z^* = 230$; autrement dit, le coût minimal de la transaction subi par Dugommier est égal au bénéfice maximal que peut réaliser Buchmoll. Ceci est un résultat plus général et caractéristique des programmes linéaires.

Reprenons le programme linéaire **(2)** ; on lui associe de semblable façon son programme dual :

Primal	*Dual*
Max $z \langle \vec{c}, \vec{x} \rangle$	Min $w \langle \vec{b}, \vec{u} \rangle$
$A\vec{x} \leqslant b$	$\vec{u}A \geqslant \vec{C}$
$\vec{x} \geqslant 0$	$\vec{u} \geqslant 0$

On montre que, à l'optimum, on a l'égalité :

(4) $$w^* = z^*$$

Cette égalité nous permet d'interpréter les variables duales $(u_1^*, u_2^*, ..., u_p^*)$, l'égalité **(4)** s'écrit :

$$\sum_{j=1}^{n} c_j x_j^* = \sum_{i=1}^{p} b_i u_i^* \quad \text{soit} \quad \frac{\partial z^*}{\partial b_i} = u_i^*$$

autrement dit, la variable duale u_i^* représente l'augmentation de bénéfice réalisée par l'accroissement d'une unité du stock de la ressource i ; u_i^* constitue donc une évaluation (interne à l'entreprise) de la ressource i ; l'intérêt économique de cette valeur est alors le suivant : soit π_i $i = 1, ..., p$, le *prix du marché* de la ressource i. Il est clair que l'entreprise aura intérêt à en acheter pour augmenter son stock tant que $u_i^* \geqslant \pi_i$. Lorsque u_i^* est nul, cela signifie que la contrainte correspondante du programme primal n'est pas saturée, c'est-à-dire que l'entreprise n'utilise pas la totalité de son stock. Toute augmentation de ce stock n'entraînerait aucun bénéfice supplémentaire.

Rôle de la programmation linéaire en économie

Les applications de la programmation linéaire sont de plus en plus nombreuses dans le domaine de l'entreprise. La programmation linéaire permet à un secteur de production de rationaliser son action en évitant le sous-emploi des ressources disponibles. Mais elle s'avère être plus qu'une technique mathématique de recherche opérationnelle : elle est avant tout un cadre conceptuel cohérent et rigoureux, très riche d'un point de vue micro-économique : l'interprétation des variables duales à l'optimum donne à la théorie marginaliste un contenu plus pertinent. En outre elle contribue à éclairer certains points de l'analyse marxiste (valeurs et prix de productions).

Concurrence et conflits - Les jeux

Dans les salons du XVII[e] siècle et dans les ruelles des grandes dames de cette époque, les jeux de société étaient fort prisés, et philosophes et mathématiciens s'essayaient à résoudre certains petits problèmes jugés fort divertissants : ainsi en témoignent les correspondances de qualité qu'échangeaient le chevalier de Méré et Pascal (1654). Mais il faudra attendre les travaux d'Émile Borel (1921) pour que ces questions reçoivent une formalisation mathématique satisfaisante. En 1928 le mathématicien allemand Johann von Neumann démontra le *théorème fondamental* ou *théorème de minimax* et joignit ses efforts à ceux de l'économiste viennois O. Morgenstern pour étudier l'application de la théorie des jeux à l'économie. Cette collaboration déboucha en 1944 sur la publication d'un ouvrage : *Theory of Games and Economic Behavior*. La théorie des jeux quittait alors les antichambres feutrées et les boudoirs fleuris pour devenir le formalisme mathématique propre à décrire les situations de conflits et de concurrence.

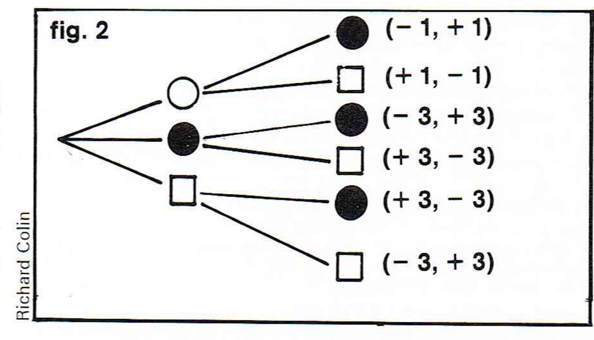
Présentation - Exemple d'un jeu à deux joueurs et à somme nulle

On appelle jeu toute « situation » dans laquelle plusieurs individus (deux ou plus) sont confrontés et doivent prendre des décisions dont dépendent des résultats. Ainsi en est-il du bridge, des échecs, du duel au pistolet, de la guerre, d'une transaction commerciale, etc. Il est clair que le résultat (ou gain) d'un joueur va dépendre non seulement de sa propre décision mais aussi de celles prises par ses protagonistes.

Exemple : deux joueurs sont face à face ; le joueur A dispose de trois jetons, deux ronds et un carré, le carré et l'un des deux ronds sont blancs, l'autre rond est noir (○ ● □) ; le joueur B ne dispose que de deux jetons, un noir rond et un blanc carré (● □). Le joueur A met un jeton sur la table sans le montrer à son adversaire qui en fait alors autant ; les jetons sont ensuite découverts. Si les deux jetons ont la même *couleur*, A donne 1 F à B, sinon c'est B qui donne 1 F à A. Si les deux jetons sont de même forme, A donne 2 F à B ; sinon c'est B qui donne 2 F à A.

Un tel jeu se représente de deux manières : une forme dite développée et une forme dite normale.

— *Forme développée*. On représente le jeu par un graphe (*fig.* 2) dont les sommets sont les points d'intervention des joueurs et les arcs les coups possibles. Pour chaque éventualité, on calcule le résultat du jeu : le premier nombre entre parenthèses indique le gain du joueur A, le second celui du joueur B.



◀ ***Figure 2 : représentation graphique du jeu (forme développée).***

— *Forme normale*. Un jeu ne se déroulant pas de manière instantanée, il y a en général plusieurs coups lorsque chaque joueur joue à son tour. En fait, on montre que tout jeu est équivalent à un jeu instantané (les joueurs ne jouent qu'une fois chacun) dans lequel une décision est un ensemble de décisions, appelé *tactique* : « Choisir une tactique revient pour un joueur à prendre globalement, avant de se mettre à jouer, toutes les décisions élémentaires qu'il peut être amené à prendre au cours du jeu » (J. Bouzitat). On représente le jeu sous la forme d'un tableau (ci-dessous) : les tactiques du joueur A correspondent aux lignes, celles de B aux colonnes. A la croisée de la ligne i et de la colonne j, on note le résultat du jeu lorsque A choisit la i-ième tactique et b la j-ième. S'agissant d'un jeu à somme nulle (ce qui est gagné par l'un est perdu par l'autre), on mentionne uniquement le gain du joueur A.

Tableau I

A \ B	●	□
○	− 1	+ 1
●	− 3	+ 3
□	+ 3	− 3

Richard Colin

Quelles tactiques vont choisir A et B ? Le choix qu'ils vont faire est commandé par la prudence : tous deux cherchent à minimiser leurs pertes. Si le joueur A joue ○, il ne peut récolter moins que « — 1 », s'il joue ● moins que « — 3 », s'il joue □ moins que « — 3 ». Il jouera donc ○. Le joueur B, quant à lui, qu'il joue ● ou □, ne peut pas perdre plus de 3. Il jouera donc, de manière équivalente, soit □, soit ●.

Le joueur A choisit donc la tactique correspondant au maximum du minimum des lignes *(maximin)*; B celle correspondant au minimum du maximum des colonnes *(minimax)*. Le maximin est égal à — 1, le minimax à + 3.

Ces valeurs sont différentes; on dit que le jeu n'a pas d'équilibre : ceci signifie que tout supplément d'information concernant les intentions d'un joueur modifiera les comportements de son adversaire : si, par exemple, B se dit « A a joué ○ », il en conclut : « je vais jouer □ ». Le joueur A, à son tour, va se dire : « je vais faire croire à B que je joue ○, ainsi celui-ci jouera □; en réalité j'aurai joué ●, ce qui me rapportera 3 F. Mais B se dira : « je devine toutes les supputations qu'échafaude A, je vais donc jouer ● et celui-ci perdra 3 F » et ainsi de suite...

Ce *jeu* est dit à *information imparfaite :* le joueur B ne connaît pas précisément ce que A a joué. Il sait seulement que le choix de A appartient à un certain ensemble = {○, ●, □}.

On peut envisager des *jeux* à *information parfaite :* à chaque coup, le joueur sait exactement ce que vient de jouer son adversaire; les échecs et les dames sont des jeux de ce type. Un théorème fondamental, le **théorème de Zermelo-von Neumann-Kuhn,** montre que tout jeu fini à information parfaite admet un équilibre. C'est dire qu'au jeu d'échecs il existe une tactique qui assure la victoire ou le match nul contre toute défense. Il serait possible de la trouver si l'on savait mettre le jeu d'échecs sous forme normale, c'est-à-dire dresser l'inventaire de toutes les tactiques possibles en fonction de toutes les éventualités envisageables par l'adversaire. Ceci dépasse largement les capacités d'énumération des ordinateurs les plus perfectionnés !

Reprenons l'exemple donné en le modifiant : le joueur A ne cache plus son jeu mais le joueur B a les yeux bandés et il peut toucher le jeton joué par A. Ce dernier est donc en mesure de distinguer deux ensembles d'information {○, ●} et {□}. Il sait si le jeton appartient au premier ensemble ou au second. Ce surcroît d'information augmente le nombre de tactiques qu'il peut envisager. Elles sont au nombre de 4 :

t_1 jouera ● si le jeton de A est rond, ● s'il est carré,
t_2 jouera ● si le jeton de A est rond, □ s'il est carré,
t_3 jouera □ si le jeton de A est rond, □ s'il est carré,
t_4 jouera □ si le jeton de A est rond, ● s'il est carré.

Sous forme normale, le jeu est représenté par un tableau à 3 lignes et 4 colonnes (ci-dessous) :

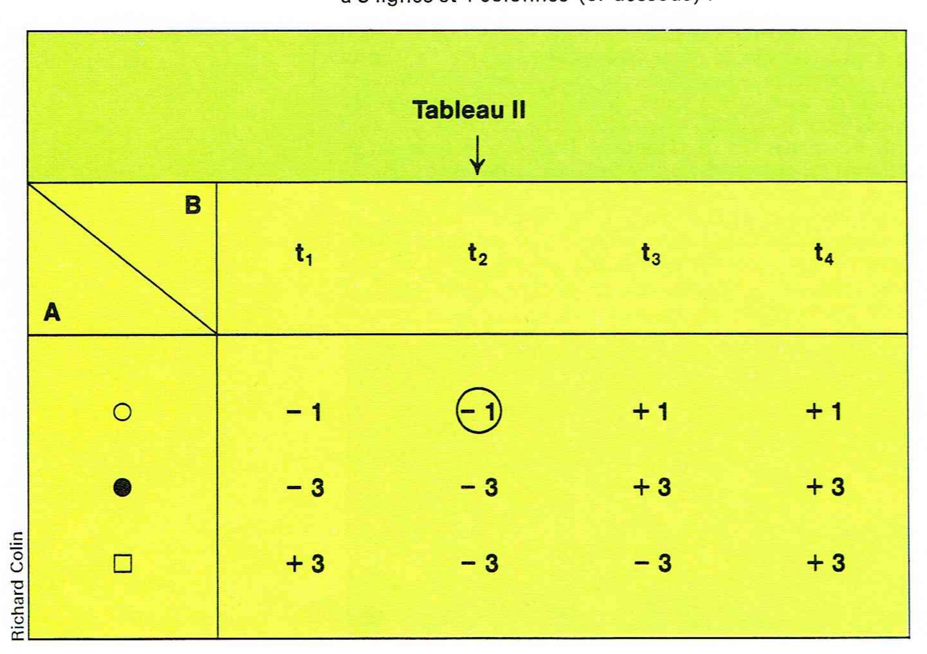


Tableau II

A \ B	t_1	t_2	t_3	t_4
○	− 1	(− 1)	+ 1	+ 1
●	− 3	− 3	+ 3	+ 3
□	+ 3	− 3	− 3	+ 3

Richard Colin

Il est clair que B va jouer t_2, car elle lui assure la perte minimale (il ne peut pas gagner moins de 1 F). Il est non moins clair que A jouera la tactique ○, correspondant à la perte minimale pour lui (il ne peut pas perdre plus de 1 F). Ici, le minimax est égal au maximin (= — 1). Les deux joueurs n'ont aucun intérêt à jouer d'autres tactiques que celles-là.

Tout supplément d'information sur les intentions d'un joueur ne modifie pas la décision de son adversaire : le joueur B ne gagnerait rien à connaître aussi la couleur du jeton de A...

Stratégies dans un jeu à deux joueurs

Un jeu à deux joueurs et à somme nulle sous forme normale est représenté par une matrice

$$A = \{a_{ij}\} \; (j = 1, ..., m; i = 1, ..., n)$$

à *m* lignes et *n* colonnes. Les lignes correspondent aux *n* tactiques du joueur (A), les colonnes aux *m* tactiques de B. a_{ij} est donc le gain réalisé par le joueur A lorsque celui-ci adopte la tactique *i* et son adversaire la tactique ***j***. En général un tel jeu n'admet pas d'équilibre, autrement dit :

$$\underset{i=1,...n}{\text{Min}} \; \underset{j=1,...m}{\text{Max}} \; a_{ij} \neq \text{Max} \; \underset{j=1,...m}{\text{Min}} \; a_{ij}$$

On montre que l'on a toujours l'inégalité

$$\text{Max Min} \leqslant \text{Min Max}.$$

Il convient donc de donner à ce concept d'équilibre une acception plus générale : on élargit la notion de tactique en définissant celle de *stratégie :* on considère alors que les joueurs A et B ont la possibilité de jouer leurs tactiques selon une certaine loi de probabilité. A chaque loi de probabilité choisie par B correspond un gain que A pourra espérer s'il choisit une tactique donnée. Ce gain espéré est défini comme la valeur moyenne des gains pour chaque tactique de B et pour la tactique donnée de A, cette moyenne étant calculée à l'aide des pondérations obtenues dans la loi de probabilité choisie par B.

Plus précisément, on appelle stratégie du joueur A une distribution de probabilité $x_1, x_2, ..., x_m$ définie sur ses *m* tactiques. Une telle stratégie sera notée X (vecteur ligne) avec $x_j \geqslant 0 \; \sum_{j=1}^{m} x_j = 1$.

De la même façon une stratégie du joueur B est une distribution de probabilité $y_1, ..., y_n$ sur ses *n* tactiques, notée Y (vecteur colonne) avec $y_i \geqslant 0; \sum_{i}^{n} y_i = 1$.

Lorsque A joue X et B joue Y, l'espérance de gain réalisée par A est donnée par :

$$\sum_{j=1}^{m} \sum_{i=1}^{n} x_j y_i a_{ij} = XAY$$

La stratégie X* optimale pour A est celle correspondant au gain maximal, une fois que B a choisi la stratégie qui défavorise le plus A :

$$X^* \text{ tel que } \underset{X}{\text{Max}} \; \underset{Y}{\text{Min}} \; XAY.$$

De même la stratégie Y* optimale pour B :

$$Y^* \text{ tel que } \underset{Y}{\text{Min}} \; \underset{X}{\text{Max}} \; XAY.$$

Le **théorème de von Neumann,** dit **théorème de minimax** (1928), montre que dans une telle situation l'on obtient toujours un équilibre (en termes de stratégie cette fois et non plus de tactique). Autrement dit :

$$X^*AY^* = \underset{X}{\text{Max}} \; \underset{Y}{\text{Min}} \; XAY = \underset{Y}{\text{Min}} \; \underset{X}{\text{Max}} \; XAY.$$

Ce résultat s'établit en remarquant que les stratégies X et Y des joueurs A et B sont respectivement solutions des deux programmes linéaires :

I	II
Max u	Min w
$XA \geqslant uJ$	$AY \leqslant wI$
$X \geqslant 0$	$Y \geqslant 0$

où $I = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \Big\} m \qquad J = \underbrace{[1, ..., 1]}_{n}$

Ces deux programmes sont en dualité. Si X* est solution de I, Y* solution de II, on a :

$$\text{Max } u = \text{Min } w = X^*AY^*.$$

Jeux à *n* joueurs

Un jeu à *n* joueurs se représente tout aussi bien sous forme extensive que sous forme normale. Sous cette dernière, on admet que les joueurs jouent de manière instantanée des tactiques $\sigma_1, \sigma_2, ..., \sigma_n$ (ou des stratégies). Chaque joueur reçoit alors un gain qui dépend de son propre choix et de celui de ses protagonistes. Mathématiquement, le joueur *i* choisit une stratégie σ_i, considérée comme appartenant à un certain ensemble Σ_i, il reçoit alors un gain $G_i(\sigma_1, \sigma_2, ..., \sigma_i, ..., \sigma_n)$. Un *n*-uplet de stratégies $\overline{\sigma}_1, \overline{\sigma}_2, ..., \overline{\sigma}_n$ sera dit en équilibre si :

$$\forall i = 1, ..., n \ \forall \sigma_i \in \Sigma_i \quad G_i(\overline{\sigma}_1, \overline{\sigma}_2, ..., \overline{\sigma}_i, \overline{\sigma}_{i+1}, ..., \overline{\sigma}_n) \leqslant G_i(\overline{\sigma}_1, \overline{\sigma}_2, ..., \overline{\sigma}_i, ..., \overline{\sigma}_n)$$

Si un tel équilibre est réalisé, aucun joueur *seul* n'a intérêt à changer de stratégie. L'existence d'un tel équilibre est assurée par certaines hypothèses mathématiques sur les ensembles Σ_i et sur les fonctions G_i **(théorème de Nash).**

Les jeux à *n* joueurs sont intéressants en ce qu'ils donnent la possibilité à certains d'entre eux de s'unir, c'est-à-dire de former des *coalitions*. Soit $N = \{0, 1, ..., n\}$ l'ensemble de tous les joueurs ; on appellera coalition toute partie S de N. Un jeu sera dit coopératif si sa règle n'interdit aucune coalition.

Considérons un jeu à somme constante :

$$\sum_{i=1}^{n} G_i(\sigma_1, ..., \sigma_n) = K \ \forall \sigma_i \in \Sigma_i$$

Supposons, pour une meilleure interprétation, que K représente un certain capital. Les joueurs appartenant à une coalition S donnée ont la possibilité d'opérer des transferts d'argent entre eux. Sous cette hypothèse, on est désormais en présence d'un duel qui oppose deux coalitions : la coalition S et la coalition N-S. Ce duel admet une solution (d'après le théorème de minimax si celui-ci est fini). Soit V (S) le gain total réalisé par la coalition S. C'est la valeur du jeu pour S : cette coalition est sûre, en se formant, de gagner au moins V (S), quoi que fassent les autres joueurs. Il est clair que l'on a l'égalité :

$$K = V(N) = V(S) + V(N - S).$$

Ainsi définit-on une application $V : P(N) \to \mathbb{R}$ que l'on appellera *fonction caractéristique* du jeu. Elle possède les deux propriétés suivantes :

(a) $V(\varnothing) = 0$

(b) $V(S \cup T) \geqslant V(S) + V(T)$ lorsque $S \cap T = \varnothing$ (l'union fait la force).

Un jeu sera dit *inessentiel* si on a l'égalité en **(b)** et *essentiel* dans le cas contraire. Dans les jeux inessentiels, les joueurs ne gagnent rien à constituer des coalitions. Seuls ceux qui sont essentiels sont dignes d'intérêt.

Une *imputation* est le résultat du jeu après redistribution des gains au sein des coalitions. Dans un jeu à somme constante il s'agit donc d'un vecteur $\vec{u} = (u_1, ..., u_n)$ tel que :

$$\sum_{i=1}^{n} u_i = V(N)$$

$V\{i\} \leqslant u_i \ \forall i$, où $\{i\}$ représente la coalition réduite au seul joueur *i*.

Cette dernière condition stipule qu'une imputation est admissible par tous les joueurs pour autant qu'elle assure à chacun au moins ce qu'il aurait gagné à jouer seul contre tous les autres.

Résoudre un tel jeu, c'est déterminer la ou les imputations correspondant à des situations d'équilibre, c'est-à-dire qu'aucun joueur n'a intérêt à refuser. On cherche à éliminer toutes les imputations qui ne pourront raisonnablement être choisies.

On dira qu'une coalition S ($\neq \varnothing$) *bloque* une imputation $\vec{u}$ si $\sum_{i \in S} u_i < V(S)$.

Quoi qu'il advienne, la coalition S refusera l'imputation *u* puisqu'elle peut gagner plus en se formant contre les autres. On en vient à définir ainsi le concept de *cœur*.



Loucel - Fotogram

▲ ***On appelle jeu toute « situation » dans laquelle plusieurs individus (deux ou plus) sont confrontés et doivent prendre des décisions dont dépendent des résultats.***

Le *cœur* d'un jeu est l'ensemble des imputations qui ne sont bloquées par aucune coalition. Il est clair que le choix d'une imputation prise en dehors du cœur conduira derechef à la formation d'une coalition qui, s'estimant lésée dans le partage, provoquera le conflit au terme duquel elle est sûre de parvenir à une position plus favorable.

Malheureusement, le cœur d'un jeu est parfois vide : on montre ainsi qu'un jeu essentiel à somme constante a un cœur vide. Ce résultat est bien troublant dans la mesure où il établit que le cœur d'un jeu coopératif à somme constante n'existe que si aucune coalition n'a intérêt à se former ! En fait les jeux à somme constante ne sont pas les plus généraux que l'on puisse imaginer (notamment dans les modèles économiques). En outre il est peu vraisemblable de supposer que toutes les coalitions ont la possibilité de se former.

BIBLIOGRAPHIE

ATTALI J., *Analyse économique de la vie politique,* P.U.F., 1972. - DESPLAS M., *Mathématique de la décision économique,* Dunod. - MUNIER B., *Jeux et Marchés,* P.U.F., 1973. - ROURE F., FRITZ G. et CHARLES A.-M., *Mathématiques pour les sciences sociales,* tome 5, P.U.F., 1973. - VAJDA S., *Théorie des jeux et Programmation linéaire,* Dunod, 1968. - ZIONTS S., *Linear and Integer Programming,* Prentice Hall, 1974.

▼ ***Le jeu de dames est dit à information parfaite.***



Keystone

MATHÉMATIQUES ET SOCIÉTÉ

Nous recourons aux mathématiques pour connaître et dominer les divers phénomènes naturels, mais aussi pour organiser toutes les informations que nous recueillons de quelque manière, encore que nous sachions que l'approche mathématique laisse échapper ou écarte délibérément maintes circonstances de ce que nous observons. Or cette approche n'a pas toujours existé et n'est pas une disposition naturelle à toute société. Cela impose une première constatation : nous vivons dans une société à laquelle l'élément mathématique est, sous une forme ou sous une autre, intimement incorporé. Non qu'il n'y ait pas eu de mathématiques ailleurs, mais c'est dans la société actuelle que leurs applications se sont étendues, que leur développement est systématiquement encouragé, que leur existence s'est vraiment institutionnalisée, bref qu'elles jouissent d'un véritable privilège social.

Cette impression peut être confirmée non seulement par des observations courantes, mais encore par une analyse historico-sociale du développement des mathématiques depuis la fin du XIXe siècle. Certes, pendant tout le XVIIIe siècle et au début du siècle suivant, les mathématiques « régnaient dans les Académies », en vertu du fait, universellement reconnu, que leur discipline avait valeur instrumentale, qu'elle était indispensable aux sciences de la nature ; mais cette prééminence était tout de même loin de s'affirmer aussi amplement qu'aujourd'hui, c'est-à-dire au point, comme on l'a écrit, de « diriger notre monde au nom de sa logique inexorable ». Au contraire, ainsi que l'a remarqué Pierre Samuel, « pendant fort longtemps, il allait presque de soi qu'un élève destiné à de longues études et à de hautes fonctions sociales, devait faire du latin et même, si possible, du grec ». Ce fait vaut d'être souligné, car il a une signification sociale ; et il est même l'indicateur principal du rôle actuellement dominant des mathématiques dans la société.

Étant donné son caractère relativement récent, cette prééminence des mathématiques doit pouvoir s'expliquer. Elle n'est sans doute pas due à une subite sensibilité collective aux charmes de cette discipline on ne peut plus abstraite, et personne aujourd'hui n'irait justifier la vogue mathématicienne, s'il y succombe, par l'attrait que cette science exerce en vertu des rapports « d'ordre, de proportion et d'harmonie » qui s'y manifestent, ou de la beauté et de la symétrie des théorèmes qui y sont établis. D'une part, personne n'est mathématicien né, et nous ne sommes plus à l'époque de Pythagore, d'autre part, les mathématiques connaissent aujourd'hui bien des phénomènes discordants, et ne s'intéressent pas aux seules propriétés où se manifestent ordre et beauté. Du reste, la tendance actuelle est plutôt, par exemple, à la préférence de l'algèbre à la géométrie, de la mémoire à l'imagination, si l'on nous permet d'employer ce langage désuet. Or, et nous citerons de nouveau un mathématicien, « l'éviction de la géométrie serait essentiellement un problème sociologique ». Il y a donc quelque chose, dans les mathématiques, qui concerne la société et se trouve concerné par elle.

La question des rapports entre les mathématiques et la société se pose donc naturellement aujourd'hui. Mais où trouver les éléments d'une réponse, sinon dans une théorie sociologique de la connaissance mathématique ? Et d'abord, y a-t-il déjà des théories sociologiques de la connaissance mathématique susceptibles de nous mettre sur la bonne voie ?

Pour étayer les remarques par lesquelles nous avons commencé, nous devrions disposer d'une théorie du savoir mathématique qui nous explique les aspects de son développement et les faits qui échappent à la contrainte strictement mathématique. Or une telle théorie n'existe pas. Bien que la sociologie de la connaissance soit née avec la sociologie même, les doctrines sociologiques n'ont manifesté aucun intérêt spécifique aux mathématiques. C'est donc par le biais d'une sociologie générale de la connaissance que, pour le moment, nous aborderons le problème.

Mais ce recours ne saurait nous satisfaire entièrement, car il peut arriver qu'une théorie sociologique décrive assez exactement des faits ou des phénomènes vrais pour certains types de connaissance, mais non vrais pour la connaissance mathématique et n'offrant rien que la communauté mathématicienne puisse reconnaître comme essentiel à son intérêt le plus proche. On peut apporter des preuves certaines d'une « collusion », ou d'une « complicité » ou d'une « correspondance » entre une société donnée et telle « philosophie », « idéologie », « théorie sociale » ou « théorie politique ». Mais le genre de parallèle qui donne généralement lieu à un rapprochement entre les produits « idéologiques » d'une société et certains de ses aspects infrastructurels — c'est-à-dire non mentaux — est peu convaincant lorsqu'il s'agit de sciences, et de mathématiques en particulier.

En effet, s'il est aisé de montrer que toute science exacte institutionnalisée véhicule un ensemble d'idées et de représentations qu'elle n'implique pas logiquement, il est quasi impossible de découvrir une homologie qui ne soit ni arbitraire ni triviale entre le statut des disciplines mathématiques d'un côté, et de l'autre les caractères par lesquels on décrit, par exemple, l'entrepreneur capitaliste ou l'idéologie technocratique. Par ailleurs, il est certain que « l'influence sociale », la mainmise de la société sur les mathématiques, se fait par le truchement d'autres disciplines, les sciences physiques le plus souvent, la philosophie parfois ; pour ce qui est des mathématiques elles-mêmes, on est en présence de structures abstraites auxquelles il est impossible de faire correspondre des segments de réalité sociale vivante, ou des segments d'idéologie. En effet, l'étalonnage socio-idéologique des formes de la mathématique permettrait de rapporter à sa place idéologique et sociale toute forme historique ou géographique de la mathématique. Ce qui ne saurait se faire sans une identification et une distinction des formes de la mathématique. Or, cette tâche, discutable dans son principe, est difficile à réaliser dans la pratique, car on ne peut faire correspondre univoquement à une formation sociale donnée une forme mathématique déterminée.

Les théories sociologiques

On n'en doit pas moins reconnaître que, malgré leur insuffisance à suivre les axes de la pensée scientifique actuelle, malgré leur vieillissement, et malgré leur défaut, d'une étude précise des mathématiques en particulier, distinctes des autres secteurs scientifiques, les théories sociologiques de la connaissance ont eu le mérite incontestable d'introduire la *problématique sociale* du savoir. Or, pour que cette problématique apparaisse, il fallait que l'élément social cessât d'être quelque chose d'accessoire et de complémentaire, imaginé par la nature pour parvenir à ses fins, comme le croyait Kant, par exemple, lorsqu'il écrivait que « toute culture, tout art formant une parure de l'humanité, ainsi que l'ordre social le plus beau, sont les fruits de l'insociabilité, qui est forcée par elle-même de se discipliner, et d'épanouir de ce fait complètement, en s'imposant un tel artifice, les germes de la nature ». Il fallait prendre au sérieux les relations de fait et de principe qui existent entre les développements du savoir et les formes de la culture, dont le savoir porte l'empreinte et dont il n'est jamais qu'une province. Il fallait substituer à l'analyse, conçue dans le style du XVIIIe siècle, des propriétés et des formes de la représentation qui permettaient la connaissance en général, la nécessité de mettre au jour les conditions de la connaissance à partir des contenus empiriques qui sont donnés en elle. Aussi est-ce chez Auguste Comte que l'on trouve une théorie du savoir qui veut en être la sociologie, et une sociologie où les mathématiques sont analysées, avec, pour l'époque, tout l'intérêt qu'elles méritent.

Le point de vue de Comte ouvre effectivement une perspective à l'étude sociale de la connaissance, parce qu'il se propose la fondation d'une science de la société, susceptible de fournir les bases d'une philosophie de la connaissance. Aux yeux de Comte, il n'y a pas de connaissance qui ne s'explique par la nature sociale des hommes et qui cesse un seul instant d'être enfermée dans quelque société.

La théorie de Saint-Simon

D'aucuns objecteraient que c'est là un type de démarche antérieur au positivisme ; que Saint-Simon, par exemple, est allé plus loin qu'Auguste Comte, dans cette voie, par sa critique vigoureuse de Condorcet qui affirmait la primauté du progrès de la connaissance sur le progrès social.

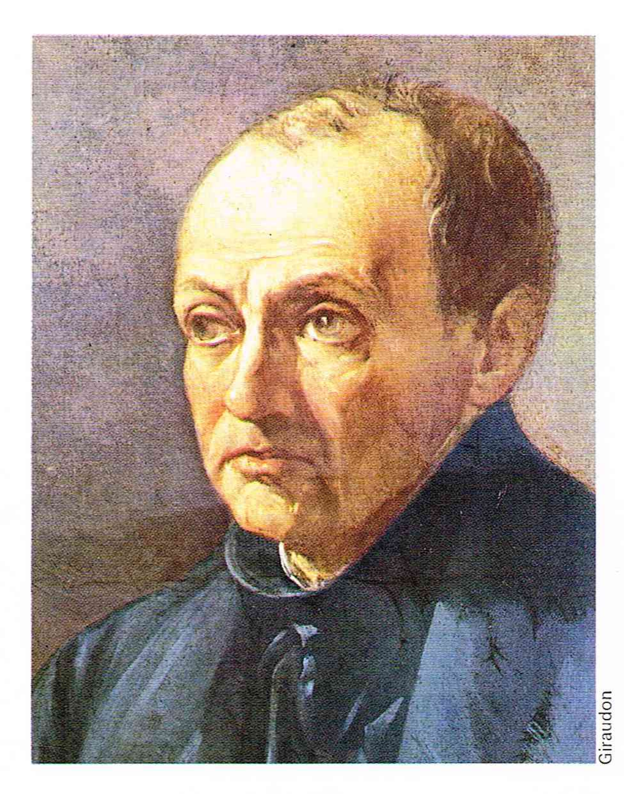
Saint-Simon fut, en effet, le premier à avancer aussi clairement l'existence d'une correspondance constante, en tous les temps, chez tous les peuples, entre les institutions sociales et les connaissances, correspondance que Condorcet saisissait globalement dans l'association naturelle, en un seul mouvement, de tous les types de progrès, dans les sciences aussi bien que dans les mœurs. D'autre part, Saint-Simon rend, certes, possible une réflexion sociale sur le savoir lui-même, dans la mesure où il marque une tendance profonde au panthéisme, laquelle permet de traiter les réalités intellectuelles et les réalités concrètes comme des émanations d'une seule nature et de considérer, ce qui est nouveau sous des mots anciens, que la « science de l'homme » étudie aussi bien « la production des biens matériels par le travail sous différentes formes » que « la production des manières de connaître ». Certes encore, tous les types de travail et toutes les manières de produire ne sont susceptibles d'une étude que parce qu'il existe « une science, bien plus importante pour la société que les connaissances physiques et mathématiques », et cette science, c'est celle qui étudie les fondements de la société elle-même. Or, cette science révèle le potentiel de la collectivité humaine dominée par l'avènement de la classe industrielle, avènement qui fait penser que « le paradis n'est pas en arrière de nous ou dans la vie céleste, mais dans notre vie future, sur terre ». Certes, enfin, tout cela ne va pas sans privilèges pour les mathématiques. Car, si « le parti industriel » est investi d'un tel espoir, c'est précisément parce que « ce parti possède la force du raisonnement », « ceux qui cultivent les sciences positives (qui sont les meilleurs raisonneurs) étant de son côté », et devenant de plus en plus nombreux.

Les transformations sociales ont, en effet, entraîné une nouvelle hiérarchie du savoir, où les sciences positives occupent la première place. « Au XVe siècle, écrit Saint-Simon dans le *Mémoire de la science de l'homme* (1813), l'enseignement public était presque entièrement théologique. Depuis la réforme de Luther jusqu'à la brillante époque du siècle de Louis XIV, l'étude des auteurs profanes, grecs et latins, s'est introduite par degrés dans l'instruction publique... de manière que la science dite sacrée a été reléguée dans des écoles spéciales, auxquelles on a donné le nom de Séminaires... Sous le règne de Louis XV, les sciences physiques et mathématiques ont commencé à faire partie de l'instruction publique ; sous le règne de Louis XVI, elles y ont joué un rôle important ; enfin, les choses sont arrivées au point qu'elles forment aujourd'hui la partie essentielle de l'enseignement... Telle est la différence, à cet égard, entre l'ancien ordre et le nouveau... que pour s'informer... si une personne avait reçu une éducation distinguée, on demandait : Possède-t-elle bien ses auteurs grecs et latins ? et qu'on demande aujourd'hui : Est-elle forte en mathématiques ? »

On pourra méditer ce texte, que nous avons cité pour son actualité. Et pourtant on ne peut véritablement parler, chez Saint-Simon, d'une sociologie dont les mathématiques constitueraient une préoccupation explicite. Saint-Simon, loin de nous fournir, fût-ce les rudiments sinon les principes d'une interrogation sur les mathématiques du point de vue de la société, ne nous livre qu'un témoignage sur la conscience que le début du XIXe siècle avait du rapport entre l'enseignement des sciences positives, des mathématiques en particulier, et l'essor industriel à ses débuts. De fait, c'est à Auguste Comte que revient la priorité dans la considération des types du savoir, en tant que constitutif de l'organisation sociale, pour autant que celle-ci est chaque fois l'expression du rapport d'un collectif humain à l'univers qui l'entoure.

Le point de vue d'A. Comte

Point remarquable, A. Comte a vu et établi que la condition nécessaire et suffisante à la constitution d'une science sociale, c'est-à-dire à l'accès de la société au statut d'objet de connaissance, consiste à prendre acte du fait qu'il est devenu impossible « de méconnaître la destination finale de l'intelligence humaine pour les études positives ». Autrement dit, tout phénomène, quelles que soient sa nature et sa complexité, comporte un aspect par lequel il est accessible à une connaissance positive. Mais, simultanément, il faut prendre acte du fait que « la philosophie positive », telle qu'on peut l'appréhender à travers la constitution du savoir astronomique, physico-mathématique, chimique, etc., n'embrasse pas tous les ordres de phénomènes. Bref, il faut encore exécuter « une grande opération scientifique » pour donner son statut au savoir dont l'objet serait la sphère des phénomènes sociaux. Ce qui ne peut se faire par simple extension du domaine de positivité déjà constitué. Entendons qu'il ne saurait y avoir de science sociale par simple transport de méthodes déjà établies ; en un mot, il ne suffit pas d'appliquer des procédés numériques aux phénomènes sociaux pour constituer une science sociale. La société, étant le lieu même où tout savoir se produit, n'est pas accessible à l'un des savoirs qu'elle constitue. C'est pourquoi toute la philosophie positive ne vise enfin, à travers la classification des sciences qu'elle a promue au niveau d'un problème philosophique fondamental, qu'à établir « la prééminence philosophique de l'esprit sociologique sur l'esprit mathématique », car « la liaison mathématique entre les phénomènes ne saurait être que précaire et stérile, en même temps que forcée et insuffisante », si l'on se contente de la fonder sur de « vagues et chimériques hypothèses ». La société, comme objet de science, est irréductible aux autres phénomènes, plus élémentaires.


Giraudon

◀ *Le philosophe français* **Auguste Comte** *(1798-1857) ; c'est à lui que revient la priorité dans la considération des types du savoir, en tant que constitutif de l'organisation sociale.*

Il en résulterait normalement que toute manifestation particulière du savoir se trouverait en liaison fondamentale, d'abord de manière générale avec les conditions organiques dont elle dépend, ensuite, et en particulier, avec la forme de société où elle se développe. La loi historique des trois états exprime cette relativisation des connaissances, chaque branche passant par trois étapes successives : l'*état théologique* ou *fictif*, où l'on distingue encore trois âges essentiels, ceux du fétichisme, du polythéisme et du monothéisme ; l'*état métaphysique* ou *abstrait ;* enfin, l'*état scientifique* ou *positif*.

Cette progression en trois degrés se voit d'abord dans la façon dont se transmettent les mathématiques. On peut penser, par exemple, qu'à l'âge théologique, l'enseignement consiste à communiquer des recettes techniques. Qu'à l'âge métaphysique, s'effectue quelque chose comme une dispersion des mathématiques parmi des calculs exacts mais partiels. Qu'à l'âge positif, enfin, il y aurait une organisation du savoir mathématique conscient désormais de son rôle et de son potentiel.

La perspective d'A. Comte n'est heureusement pas aussi simple. A. Comte remarque d'abord qu'alors qu'un géomètre de l'Antiquité se formait par l'étude successive d'un très petit nombre de traités originaux, essentiellement les écrits d'Archimède et d'Apollonius, « un géomètre moderne a communément terminé son éducation, sans avoir lu un seul ouvrage original, excepté relative-

ment aux découvertes les plus récentes, qu'on ne peut connaître que par ce moyen ». Ce qui veut dire qu'une sociologie moderne des mathématiques est obligatoirement, pour une part au moins, une sociologie des *manuels* plutôt que des *mémoires originaux* de mathématiques. Le manuel s'est désormais imposé comme intermédiaire indispensable entre la société et les mathématiques. Ensuite, Comte saisit clairement que la révolution cartésienne, qui rendit possible la géométrie analytique, implique que l'organisation vers l'objet mathématique se soit elle-même modifiée; la révolution cartésienne est une véritable libération des forces intellectuelles, qu'elle dirige vers des *questions plus générales,* à l'aide de notions nouvelles, qu'on n'aurait jamais pu introduire « en ajoutant quelques nouvelles courbes » au petit nombre de celles qu'on avait déjà étudiées par l'ancienne méthode. Si celles-ci n'intéressent plus le mathématicien, cela tient à ce que la révolution philosophique opérée en géométrie par Descartes « a dû singulièrement diminuer l'importance de semblables recherches ».

Ne relevons pas davantage de remarques de détail, bien qu'elles soient, chez A. Comte, plus importantes que ce qu'on lit dans les développements consacrés explicitement aux mathématiques, lesquels véhiculaient les idées courantes dans le milieu de l'École polytechnique, au temps où notre auteur en était l'élève. En multipliant et en rassemblant des notations du genre de celles dont nous avons reproduit l'essentiel au paragraphe précédent, on s'apercevrait qu'elles tendent à une certaine unité et qu'elles résistent encore aujourd'hui à la critique.

Maints sociologues du savoir ne font actuellement qu'énoncer doctement des faits qui constituaient des évidences pour A. Comte, bien qu'en raison de l'idéalisme alors dominant ils ne fussent pas immédiatement intelligibles à son époque : par exemple, l'idée que l' « élément scientifique » entretient des relations, non seulement plus ou moins directes, mais encore absolument organiques, avec certaines demandes militaires. Jusqu'alors, dit Comte, la science avait reçu des encouragements facultatifs; à présent la protection des sciences devient systématique, et, « pour tous les gouvernements occidentaux, un véritable devoir », en vertu de la liaison étroite des sciences exactes avec les procédés militaires et avec l'essor industriel. Or, ce dernier est, à son tour, conditionné « spécialement par l'institution des armées soldées ». D'autres institutions, au dessein non moins militaire, semblent, d'après un mathématicien contemporain, René Thom, à l'origine de l'engouement actuel pour les mathématiques; le lancement des satellites aurait joué le rôle de l'aimant attirant l'attention du public sur les techniques mathématiques.

Il va sans dire que Comte n'a pas systématisé cette théorie des rapports entre les mathématiques et l'organisation sociale; elle reste pour une grande part implicite et ne reçoit pas toute la précision voulue. Mais justement, cette esquisse qu'il nous a laissée est d'autant plus précieuse qu'elle n'a été ni reprise ni complétée après lui. C'est qu'en fait les choses sont délicates. Comte, comme tous les philosophes traditionnels, n'a pas manqué de souligner l'originalité des mathématiques, leur autonomie (relative) à l'égard du social. Ce qui s'explique, selon lui, de deux façons. D'une part, l'esprit positif se manifeste dès l'origine, bien qu'il soit encore dominé par les croyances non positives. Le germe de la philosophie positive est tout aussi primitif que celui de la philosophie théologique, bien qu'il n'ait pu se développer que plus tardivement. D'autre part, et plus particulièrement, les rudiments de positivité sont d'autant plus anciens qu'ils sont plus simples, d'autant plus résistants à la théologie qu'ils ont un caractère de généralité et d'abstraction plus grandes. Chacun reconnaîtra ici que ce sont les idées mathématiques qui sont appelées à former le berceau de la positivité.

Bien plus, leur présence, dans la mesure où il s'agit pour Comte de mathématiques essentiellement appliquées, constituant ainsi, dans le cas de la géométrie surtout, le mode le plus simple de l'existence inorganique, marque l'introduction d'un modificateur graduel de la philosophie primitive, d'un principe de dissolution pour les croyances théologiques ou métaphysiques, sûr, efficace, mais dont l'action est, pour ainsi dire, plutôt intermittente que continue, justement parce qu'il n'est pas toujours accordé aux conditions sociales qui l'entourent.

Avec Thalès, Pythagore, Hipparque et Archimède — Archimède surtout, dont l'intérêt pour les applications pratiques préfigure l'idée des immenses services que la science devait rendre un jour à l'industrie — l'esprit scientifique s'était développé avec une telle plénitude qu'il suscite pour la philosophie de l'histoire une question caractéristique : comment expliquer que les dispositions éveillées n'aient pas été immédiatement mises à profit et qu'un intervalle de quinze siècles sépare l'élaboration astronomique d'Hipparque des découvertes de Kepler, pour ne prendre que cet exemple ? Ni les situations historiques, ni les mérites personnels ne renferment l'explication. La recherche de celle-ci nous renvoie à la théorie complexe des facteurs dominants, celle de la présidence, de la souveraineté, de l'ascendant de certains caractères sociaux. L'évolution scientifique dépend de tout l'essor mental, et celui-ci peut comporter des éléments incompatibles entre eux qui obscurcissent l'horizon de la science. Un horizon jamais disponible à l'état pur, multiplement connexe, multiplement compact, suffisamment mobile pour ruiner toute prétention à l'absolu, sans jamais assujettir ce qui est scientifiquement établi à une histoire purement arbitraire. Car le progrès est toujours maturité et développement de l'expérience, habité dès l'origine par des lois logiques, essentiellement invariables, communes à tous les temps et tous les lieux, et même aux sujets quelconques.

L'apport des sociologies de la connaissance

Les travaux des sociologues classiques qui se sont intéressés aux formes de la connaissance restent, pour ce qui concerne les mathématiques, bien en deçà de la profondeur des vues de Comte. Ainsi, le philosophe Jérusalem, qui a créé l'expression « sociologie du savoir » et plaidé pour une étroite dépendance du social et du pensable, n'a fait que traduire en termes sociologiques le langage transcendantal.

Le grand Durkheim lui-même est, à certains égards, en retrait par rapport à Comte; en effet, il fait dériver la connaissance de la religion, perspective hasardeuse et en tout cas moins prudente que celle de Comte qui montre une présence simultanée, dès l'origine, d'éléments positifs et d'éléments antagonistes provisoirement dominants, évitant par là l'écueil d'avoir à dériver le savoir du social ou d'un de ses aspects. Par ailleurs, Durkheim, en suggérant la relation des catégories d'espace et de temps avec la structure sociale où elles sont produites, veut assigner une origine sociale à l'ensemble des catégories de l'entendement, c'est-à-dire aux concepts de genre, de nombre, de cause, de substance, par quoi il se rend tributaire de la théorie de la connaissance de Hamelin. Aussi sa sociologie véhicule-t-elle, quand elle se passe d'enquête empirique, les préjugés typiques d'une *théorie de la connaissance* bien datée. Enfin, et surtout, les mathématiques n'intéressent ni directement ni explicitement Durkheim, car cela n'avance à rien de dire que l'idée de nombre correspond aux propriétés les plus universelles de l'être.

On pourrait attendre mieux de Lucien Lévy-Bruhl. Celui-ci, en effet, était plus sensible à la pluralité des genres de connaissance, et, dans sa dernière œuvre, à la diversité de leurs rôles dans différents types de société. On sait qu'il a d'abord soutenu la thèse de la loi de « participation mystique » comme substitut, dans les sociétés archaïques, des principes logiques ayant cours dans les sociétés « civilisées ». Par la suite, il a évolué vers l'idée d'une logique primitive spécifique, opposée à la logique développée sous la contrainte de la rationalité. Mais nulle part notre problème n'est envisagé pour lui-même, ni ne bénéficie des résultats recueillis. C'est pourquoi les autres efforts peuvent paraître également décevants. Une exception cependant : les analyses de Max Scheler (1874-1928).

La contribution de Max Scheler

La connaissance mathématique est traitée sous la rubrique générale de la connaissance scientifique; Scheler prend, le premier, ses distances à l'égard de la philosophie de la connaissance traditionnelle. Très proche du perspectivisme de Husserl, il considère, par exemple, que la table des catégories de Kant correspond simplement à la configuration conceptuelle qui a dominé l'Europe occidentale à un certain moment de son histoire. Il relativise tous les *a priori* subjectifs, c'est-à-dire tous les « sujets

collectifs » qu'on imaginait au principe de la connaissance, échappant ainsi à la tentation de sociologiser le *cogito*, c'est-à-dire de substituer au sujet cartésien sur lequel se fonde le rationalisme classique un sujet collectif, le groupe et l'époque. L'intérêt principal de sa sociologie est l'analyse explicite de ce qu'il considère comme déterminant l'essence de la science positive, laquelle doit son origine à la jonction de deux couches sociales séparées au départ mais condamnées à collaborer de plus en plus rationnellement : celle des contemplatifs libres et celle des hommes de métier et d'expérience. C'est la première qui garantit à la seconde le fondement méthodique, logique et mathématique ; la seconde procure le rapport à la technique, à la mesure, à une expérimentation d'un type nouveau, échappant au hasard, impliquant une manipulation des corps et des forces naturelles familière en premier lieu aux cultures patriarcales expansives, à des sociétés organisées différemment des sociétés matriarcales, introverties.

L'attitude de Scheler consiste donc à refuser à la fois le pragmatisme, la conception de la science positive ou des mathématiques en particulier, qui en fait une science totalement issue de la technique, et la conception intellectualiste qui n'entend la relier qu'à la philosophie. Partout où il y eut science positive, c'est donc sur la base de l'association de la philosophie avec l'expérience du travail. Les formes des techniques de production et du travail humain correspondent aux formes de la pensée scientifique et positive sans que les unes soient les causes ou les effets des autres. La technique n'est pas la simple application d'une science pure qui la précède ; car la volonté de puissance dirigée sur tel ou tel domaine détermine à l'avance les méthodes de pensée et d'intuition aussi bien que les buts de l'activité scientifique, mais elle les détermine de manière inaccessible à la conscience des individus, hors du champ de visibilité immédiate du chercheur ou du savant.

Il y aurait donc lieu d'établir une typologie générale des systèmes de pensée qui se succèdent en reprenant chacun (et chacun à sa façon) des problèmes principaux qui se posent. Mais par là Scheler tend à se rapprocher de la théorie des visions du monde, chère à Dilthey, et trop compacte pour nous apprendre quelque chose de précis sur notre sujet.

En effet, dire avec Scheler que la « volonté de puissance » habite la profondeur invisible du projet scientifique rend la mathématique elle-même, si pure qu'elle soit, à la réalité sociale où elle se forme : l' « idéologie » habite fondamentalement le savoir. Mais c'est là une vérité qui ne nous apparaît qu'à travers des analyses conceptuelles parfois trop abstraites pour être concrétisées, ou trop engagées dans la philosophie pour garder toute leur signification sociale. C'est sans doute la raison pour laquelle on ne trouve, en fin de compte, aucune analyse concrète des rapports entre mathématiques et société dans les travaux de Scheler, ni, non plus, dans ceux de Karl Mannheim, celui qui fut lesociologue de la connaissance par excellence.

L'analyse de Karl Mannheim

En effet, Karl Mannheim, parti d'une réflexion sur l'interprétation des œuvres culturelles, sur l'analyse structurale de la théorie de la connaissance et sur la problématique d'une logique de la philosophie, a développé sa théorie de la connaissance d'abord dans le cadre des problèmes inhérents à l'interprétation sociologique des œuvres, avec le souci de voir si l'examen peut mettre en lumière dans l'œuvre quelque chose qui échappe à l'ensemble des conditions sociales. Mannheim aperçoit la distance qui sépare l'individu de l'œuvre objective, qu'elle soit artistique ou scientifique. En France, L. Goldmann et même J.-P. Sartre retrouvent la même réflexion sur la notion d'œuvre, mais ne considèrent guère plus que Mannheim les mathématiques en particulier, peut-être parce que là, plus qu'ailleurs, apparaît la fragilité de cette notion d'œuvre. Par ailleurs, Mannheim s'en tient à des questions de nature trop existentielle : Où sommes-nous ? Qu'est-ce qu'interpréter ? Quelles sont les différentes formes de l'interprétation ? ou trop générale : Que veut dire « connaître » ? Où en sommes-nous du devenir historique ? De quel lieu envisageons-nous notre univers intellectuel ? questions que leur résonance pathétique situe à leur tour tout à fait historiquement et socialement.

L'analyse marxiste

La philosophie marxiste, de Marx à ses interprètes récents, est restée plus fidèle à l'inspiration concrète, à l'analyse du procès de la connaissance réelle, non seulement en tant que fonction d'un ensemble de conditions historiques dont on définit la nature et les articulations, mais encore et surtout en tant que travail actuel élaboré sur la base des résultats de travaux effectivement disponibles. Or, par ce biais, le marxisme tendait plutôt à élaborer une épistémologie concrète des mathématiques : la dialectique interne d'un mouvement singulier en tant qu'il est révélé, strictement parlant, par ce mouvement même. Une certaine autonomie du secteur mathématique s'affirme par là, et contredit la thèse générale d'Engels, pour qui la science est l'une des régions idéologiques supérieures, le besoin économique ayant été et n'ayant cessé de devenir davantage le « principal ressort du progrès de la connaissance de la nature » ; les scientifiques s'imaginent seulement « qu'ils travaillent sur un terrain indépendant » ; en réalité, tout en constituant un groupe autonome au sein de la division sociale du travail, tout en exerçant, en retour, une certaine influence sur le développement social, voire le développement économique, ils n'en sont pas moins sous l'influence dominante de ce dernier. Cependant, toute la difficulté consiste à interpréter, conceptualiser l'idée de cette « influence » qui, pour éviter toute connotation mystique, devra être dissociée en la possibilité de l'intervention des résultats scientifiques dans tel domaine économique et social, et, d'autre part, en pratique relativement autonome, avec une structure et une histoire propres.

Telles sont les théories qui envisagent plus ou moins implicitement le problème des rapports entre les mathématiques et la société. Nous n'avons point considéré celles qui n'en traitent que de manière fort indirecte : celle de Max Weber, par exemple, sous la forme de la rationalité occidentale comme effet, au même titre que le capitalisme moderne, de l'éthique protestante. Et nous avons rencontré un certain nombre de questions qui auraient pu donner lieu à des enquêtes intéressantes sur les rapports des mathématiques et de la société aujourd'hui, mais toutes les théories dont nous avons parlé restent épistémologiquement en retard sur les mathématiques. Les indications que nous avons recueillies constituent donc les cases vides d'un programme inaccompli.

L'impact mathématique sur le social

Une réflexion plus actuelle s'avère donc indispensable. Aujourd'hui, en effet, la mathématique semble être devenue une préoccupation éminemment sociale, et de plusieurs manières :

— en tant qu'elle est devenue un outil indispensable, bien qu'insuffisant, à toute science sociale ;

— en tant que mode de savoir dominant de la société technologique ;

— en tant, enfin, qu'elle s'est imposée comme constituant essentiel d'une nouvelle culture de base.

Tout d'abord, la mathématique a conquis, il est vrai, le domaine social. La sociologie, par exemple, ne recourt pas seulement aux méthodes quantitatives pour l'analyse des données, mais à une foule d'outils mathématiques plus perfectionnés.

Elle a ses origines propres. Elle ne semble pas résulter de la simple extension au social des méthodes qui ont fait leurs preuves dans les domaines de la nature. C'est pourquoi l'on ne trouvera pas à l'origine du projet d'une science sociale le souci de construire un *concept abstrait* de la « société ». Autrement dit, si c'est la sociologie empirique qui s'est le plus ouverte, et le plus utilement, aux mathématiques, le projet de cette sociologie empirique existe tel quel dès l'apparition des premiers efforts pour comprendre les divers phénomènes aujourd'hui rangés sous la rubrique des sciences sociales.

Histoire et valorisation idéologique de la mathématisation du social

Les premiers essais pour introduire la quantification dans les sujets d'ordre social coïncident avec la naissance de la science moderne du mouvement. Le même siècle a enfanté deux disciplines aussi différentes que la mécanique et la sociologie quantitative. Certes, c'est, ici et là, le même besoin de maîtriser l'apparence désordonnée

en fondant une approche méthodique et réfléchie, d'une part, des phénomènes naturels, de l'autre, des phénomènes sociaux. Mais chacune de ces deux entreprises, qui n'eurent d'ailleurs pas un égal succès, mettait en œuvre une rationalité manifestement hétérogène à celle de l'autre. L'outil de la physique mathématique n'est pas celui qui sert à déterminer si un ensemble de données numériques, recueillies d'une certaine façon, permet de dégager une certaine régularité, sinon d'établir une nouvelle espèce de loi.

La coïncidence historique de ces préoccupations d'orientation opposée décèle une intention de rationalité, qu'il serait peut-être imprudent de rapporter, comme à leur cause commune, à l'esprit du rationalisme montant. Il se peut bien que cette rencontre ne fût pas l'œuvre d'un pur hasard mais, actuellement, aucune recherche ne permet d'en établir la nécessité de manière convaincante.

Cependant, quelques faits bien établis révèlent certaines dépendances entre les préoccupations de rationalité économique et sociale et certaines exigences mercantilistes : la création des systèmes d'assurances, l'organisation du crédit public, le recours aux dénombrements démographiques, etc., toutes choses rendant indispensable un fondement numérique sûr et permettant d'établir un ordre financier rigoureux, de limiter le gaspillage, de réduire le parasitisme social favorisé par les modes de vie plus anciens, de valoriser, enfin, le travail en y soumettant les princes eux-mêmes dont les fonctions devaient désormais être liées à des objectifs économiques précis.

La nouvelle dimension des États et leur inféodation à la conquête des premières richesses exigeaient donc une meilleure connaissance de leur situation démographique et financière, et ont, par conséquent, directement commandé le développement de la *statistique,* comme science de l'État (Staatskunde), ainsi que le suggère le mot lui-même, introduit en 1749 par Achenwall, professeur à Göttingen. Certes, en tant que simple désignation de l'activité qui consiste à recueillir des données permettant de connaître la situation des États, les statistiques trouvent un ancêtre dans « la science des listes » si développée par les Sumériens qui ne négligeaient jamais de répartir par séries et par catégories les données de l'expérience, et plus encore, dans le recensement des productions agricoles qui remontent, en Égypte, à plus de 2 000 ans avant Jésus-Christ : les pharaons cherchaient, en effet, le moyen infaillible d'obliger tous les chefs de famille à payer les taxes individuelles et personnelles ; plus récemment, le capitalisme des grands marchands de drap de Venise se dota d'une organisation rationnelle et d'une structure bureaucratique, qui a pu fournir un modèle aux Pays-Bas du XVIIe siècle, où la renaissance du commerce s'est accompagnée d'un renouveau de l'enquête statistique attesté par les quelque soixante volumes des *Respublica alzeviriana,* où l'on trouve des informations sur l'économie des États.

C'est bien au XVIIe siècle que l'on peut trouver l'ancêtre direct de la statistique actuelle. Car c'est au XVIIe siècle que s'accuse le plus nettement le rapport profond entre la société et l'État. La définition de la statistique que retient Cournot, dans sa *Théorie des chances et des probabilités,* nous montre rétrospectivement la conscience qu'on avait de ce rapport : « On entend principalement, nous dit-il, par statistique, comme l'indique l'étymologie, le recueil des faits auxquels donne lieu l'agglomération des hommes en sociétés politiques... en tant que ces faits sont susceptibles de dénombrement et d'évaluation numérique. »

C'est Colbert, en France, qui avait ordonné des enquêtes fournissant toutes ces statistiques dont regorgent les mémoires des intendants. Le procédé devient systématique, et comme un moyen de gouvernement. Et graduellement, la perception du socio-politique devient aussi peu commune et immédiate que celle des phénomènes naturels. Si bien qu'on peut bientôt dire des phénomènes socio-culturels ce que Newton disait des phénomènes naturels : « Le moindre fait qui s'offre à nos yeux est tel qu'on ne peut sans une extrême adresse démêler tout ce qui y entre, ni même sans une sagacité extrême soupçonner tout ce qui y peut entrer. »

Mais si le lieu d'application originaire de la statistique est le domaine socio-politique, c'est d'avoir fonctionné comme *lieu d'application de la théorie des probabilités* qu'elle a eu un destin tout autre que celui des divers recueils de données sur la situation des États.

Dans cette perspective, où les statistiques sont liées à des considérations probabilistes, on notera que, dès 1570, Cardan s'était intéressé aux statistiques relatives à la durée de la vie humaine ; en 1693, l'astronome anglais Halley n'a pas dédaigné de publier une *Breslau Table of Mortality,* première tentative d'établir une table des mortalités sur des données concrètes. Ces tables de mortalité furent un des premiers résultats à répandre le culte des déterminations numériques dans l'ordre socio-politique, comme le montre l'œuvre de Süssmilch (1707-1767) sur *l'Ordre divin prouvé par la natalité, la mortalité et la fertilité du genre humain.* L'engouement devient bientôt général : y participent non seulement les théologiens mais aussi les hommes de lettres et de sciences, et Buffon, par exemple, le justifie en alléguant que « de toutes les probabilités morales possibles, celle qui affecte le plus l'homme en général, c'est la crainte de la mort ». En fait, il est la marque du succès des premières statistiques médicales, commencées dès l'œuvre de Graunt (*Natural and Political Observations Mentioned in a Following Index and Made upon the Bills of Mortality,* 1662), et se poursuivant tout au long du XVIIIe siècle, où apparaît justement dans la littérature la nouvelle rhétorique toute logique et toute rationnelle, fortement imprégnée du prestige des chiffres.

Qu'on en juge par ce discours de Voltaire sur la prophylaxie de la variole : « Il y a quelques gens qui prétendent que les Circassiens prirent autrefois cette coutume des Arabes ; mais nous laissons ce point d'histoire à éclaircir par quelque savant Bénédictin, qui ne manquera pas de composer là-dessus plusieurs volumes *in-folio* avec les preuves. Tout ce que j'ai à dire sur cette matière, c'est que, dans le commencement du règne de Georges Ier, Mme de Wortley-Montaigu... avec son mari en ambassade à Constantinople, s'avisa de donner sans scrupule la petite vérole à un enfant dont elle était accouchée en ce pays... Cette dame, de retour à Londres, fit part de son expérience à la princesse de Galles, qui est aujourd'hui reine... Dès qu'elle [la reine] eut entendu parler de l'inoculation ou insertion de la petite vérole, elle en fit faire l'épreuve sur quatre criminels condamnés à mort... »

Suit l'argument fondamental aux yeux de Voltaire : « Sur cent personnes dans le monde, soixante au moins ont la petite vérole ; de ces soixante, vingt en meurent dans les années les plus favorables et vingt en conservent pour toujours de fâcheux restes : voilà donc la cinquième partie des hommes que cette maladie tue ou enlaidit sûrement. »

La leçon à en tirer est qu' « une nation commerçante est toujours fort alerte sur ses intérêts, et ne néglige rien des connaissances qui peuvent être utiles à son négoce » *(cf. la onzième lettre philosophique, sur l'insertion de la petite vérole).*

Voilà pourquoi Diderot, de son côté, déplore qu'on ait donné trop d'importance dans les écoles à l'étude des mots, alors que le spectacle de l'industrie humaine est en lui-même grand et satisfaisant pour développer « l'instinct de la précision », pour faire sentir, dans les cas de *probabilité,* l'écart plus ou moins grand par rapport au vrai, pour faire apprécier les incertitudes, « calculer les chances », faire sa part au sort ; bref, c'est en ce sens, conclut Diderot, que les mathématiques deviennent une science usuelle, une règle de la vie, une balance universelle.

Le talent littéraire s'est donc lui-même mêlé de cette valorisation du nombre, de la quantité et de la statistique : d'où cet engouement, caractéristique du XVIIIe siècle, bien exprimé par Voltaire, ambitieux de réunir le titre de géomètre à celui de poète et d'historien ; d'où l'admiration de tous pour Buffon, qui communique aux sciences, y compris l'arithmétique morale ou politique, le charme dont les lettres avaient eu jusque-là l'exclusivité. Il en résulte une certaine popularité des sciences en général, assez profonde en France où les meilleurs ouvrages sont écrits dans un style à la fois accessible et modèle, lors même que les ouvrages de Gauss ou de Newton demeuraient écrits en latin.

Cette pénétration de la littérature elle-même par la science et par ses applications marqua le point de départ d'une période qui vient à peine de s'achever, et dont la fin est saluée par les mathématiciens contemporains qui



◀ *Le penseur prisonnier des algorithmes (vu par D. Ribas).*

D. Ribas

se félicitent que les mathématiques puissent aujourd'hui s'exposer avec une rigueur telle qu'elle exclut le genre d'« exposé décoratif » qui permettait à certains de briguer à la fois l'Académie des sciences et l'Académie française ! C'est que ce prestige des mathématiques s'illustrait, en effet, plus par une précision excessive à laquelle se joignait la passion du pittoresque, que par un souci d'objectivité et de rigueur. La mathématique offrait un arsenal de métaphores et de modèles pour concrétiser telle ou telle idée. Même Rousseau n'échappe pas à cette façon de faire, qui expose dans le *Contrat social* une théorie du gouvernement sur le modèle d'une théorie des proportions.

Le désir d'étendre les mathématiques aux domaines appartenant à la société plutôt qu'à la nature existe donc réellement, surtout chez les premiers statisticiens. Mais ces efforts précurseurs ne dépassent pas le stade préscientifique : les enquêtes statistiques de ce temps respectent rarement les normes de l'enquête empirique telles qu'elles seront dressées plus tard, et suppléent le vide de l'information par l'audace des déductions. C'est pourquoi l'histoire de la *quantification* du social ne doit pas négliger les oppositions à la statistique. Celles-ci ont été exprimées au cœur du XIXe siècle par A. Comte et réitérées par Claude Bernard, qui est sans doute allé plus loin dans la critique et la mise en valeur du ridicule des statistiques aveuglément appliquées, par exemple, dans le cas où l'on recueille l'urine d'un homme pendant vingt-quatre heures et qu'on mélange tous les échantillons pour avoir l'analyse de « l'urine moyenne » (on aura en effet l'analyse d'une urine qui n'existe pas). Ce qui a rendu la statistique inacceptable aux esprits les plus positifs, c'est précisément sa tendance à chiffrer n'importe comment, à quantifier sans méthode, si bien que Cournot, auteur du premier effort sérieux de mathématiser la théorie des richesses, pensait qu'un statisticien ou un financier ne peuvent guère être considérés comme des mathématiciens, et que nombres et mesures n'étaient pas forcément des mathématiques. C'est dire que *le besoin de gestion ou d'administration ne peut tenir lieu de fondement scientifique, ni de concept, ni de théorie, ni de légitimation philosophique, et si l'application des statistiques au social a peu à peu balayé les scrupules, c'est que le perfectionnement de la science a induit un raffinement de l'analyse des faits considérés.*

Amélioration des méthodes statistiques et utilisation de méthodes mathématiques non statistiques

L'engouement pour les statistiques ne fera plus sourire à partir du moment où elles seront fondées sur le calcul des probabilités. Or, Adolphe Quételet (1796-1874) fut, à coup sûr, le premier à voir tout le parti que pouvait tirer une « physique sociale » des travaux de Fourier et de Laplace sur la probabilité mathématique. D'abord intéressé par des moyennes et des taux pour des propriétés généralement d'ordre démographique, il étendit bientôt ses investigations, vers 1840, à la *distribution* de ces propriétés, en se laissant guider par l'analogie qu'il remarquait entre la distribution de la taille et du poids des êtres humains et celle, mieux connue, des erreurs d'observation. C'est là une application anthropologique qui inspire encore aujourd'hui l'histoire quantitative lorsqu'elle étudie, par exemple, les archives de recrutement militaire.

Mais l'œuvre de Quételet restant dans l'ensemble assez confuse, et encore implicite quant à ses principes, suscitera des critiques, dont les plus célèbres sont celles que lui adressa M. Halbwachs dans *la Théorie de l'homme moyen* (1912). Lazarsfeld remarque à juste titre qu'en réalité c'est seulement après plusieurs décennies, avec le développement des « processus stochastiques », que le mouvement inauguré par Quételet fut repris et que fut démontrée l'applicabilité du calcul des probabilités à ce que Halbwachs appelait le domaine de « l'interaction sociale ».

Par exemple, la distribution binominale exprime la situation d'un bal où les hommes et les femmes, celles-ci étant en plus grand nombre, ne se connaissent pas : à l'ouverture du bal, chaque cavalier choisit une cavalière au hasard en tirant son nom au sort, et au bout de dix danses on peut ainsi classer les noms des danseuses selon le nombre de fois où elles ont été invitées. Or, si les cavalières choisies la première fois bénéficient de la croyance d'être les plus désirables, elles auront plus de chances d'être choisies, si bien que tout se passe comme si, dès la seconde danse, les noms des cavalières choisies la première fois étaient mis deux fois dans le chapeau. Si l'on répète l'expérience, le nombre des danseuses ayant eu le plus ou le moins de cavaliers croît sensiblement, tandis que le nombre moyen de succès demeure le même. Ainsi rien n'empêche le développement de processus stochastiques où les probabilités des choix individuels à l'instant $t + 1$ dépendent de la distribution totale de la probabilité à l'instant t. On ne peut en conclure, comme le fait Lazarsfeld, que Quételet avait raison contre Durkheim ; l'acquis positif de chacun des deux points de vue semble aujourd'hui conservé, car il n'est pas dit que l'explicitation de la méthode de Durkheim dans son étude sur le suicide ne mène pas au même résultat que la systématisation des remarques de Quételet postulant qu'on peut mesurer l'inobservable si l'on suppose des relations mathématiques entre caractères observables et variables non observables, ou notant qu'on peut substituer une seule observation dans le temps portant sur une grande quantité de gens à des observations répétées sur une même personne.

Il a fallu attendre le début de ce siècle pour disposer d'une bonne méthodologie statistique, c'est-à-dire de la théorie qui permet, grâce à un concept convenable de l'*inférence statistique,* de passer des données observables à des conclusions sur les lois de probabilité qui régissent ces données. Autrement dit, il a fallu le développement de la statistique mathématique pour que le langage des sciences sociales lui-même puisse être remis en chantier, être passé au crible d'une conceptualisation critique et d'une redéfinition de ses notions opératoires. Ce dont témoigne bien la nécessité, vivement ressentie par Lazarsfeld, d'analyser et de délimiter les objets empiriques, de clarifier et d'expliquer les termes, tâche de nature bien peu empirique, bien qu'elle soit inscrite dans le cadre d'une élucidation aussi précise que possible de l'enquête empirique. Celle-ci ne veut plus rester aveugle sur la jonction qu'elle effectue entre le langage dans lequel elle s'exprime, sur ses objets et les moyens qui permettent l'expression quantifiée. Tentative d'homogénéiser l'objet de l'enquête et son appréhension scientifique, de fixer les conditions de l'applicabilité des concepts retenus à certains ensembles d'éléments observables. Le caractère empirique des recherches de Lazarsfeld vient de la seule nécessité de repenser pour chaque cas une épistémologie locale. De fait, l'ensemble des objets sociologiques est considéré comme une *combinaison de propriétés élémentaires de variables;* un peu comme les corps sont réduits par Descartes à un espace homogène. Les relations entre variables sont une manière de monnayer, de développer les idées les plus complexes. *Variable* est, par exemple, la taille d'une ville, l'état financier d'une entreprise, le quotient intellectuel d'une personne, bref, tout ce qui se laisse mesurer d'une certaine façon. Il va sans dire que tout ne se traduit pas en « variable », avec toute la précision voulue. Mais cela permet des distinctions entre catégories de concepts, c'est-à-dire critères de classification donnés. Ainsi comprise, cette conceptualisation conforme pour ainsi dire les données des sciences du comportement, et donc des sciences sociales, aux exigences de la statistique mathématique.

La construction de modèles statistiques, qui permettent une analyse de type causal, n'est pourtant pas l'unique voie empruntée par la mathématisation du social. Bien des situations réelles relèvent de modèles mathématiques plus complexes, par exemple les *modèles stochastiques,* plus propres à rendre compte des *évolutions* d'une grandeur dans le temps, le hasard intervenant à chaque instant. Mais, dans l'état actuel du calcul des probabilités, les applications de la théorie des processus stochastiques au social restent assez limitées : l'analyse d'un modèle stochastique devient, en effet, très complexe, dès que les hypothèses qu'il exprime dépassent un niveau simple; en particulier dès que le nombre de variables intervenant dans le modèle dépasse un certain ordre, assez petit.

Bien que les statistiques aient permis de traiter mathématiquement des notions que d'aucuns croyaient réfractaires par nature, elles ne constituent certes pas le seul traitement mathématique possible des phénomènes sociaux. On connaît ainsi un exemple d'application des probabilités ayant un fondement étranger à toute préoccupation statistique. Condorcet a créé, en effet, une science sociale mathématique sur la base d'une philosophie sociale contractualiste, fondée sur les notions de volonté et d'intérêt généraux. Cette idée, qui est aussi celle de Locke, de Rousseau, de Diderot, implique une conception de la conduite humaine en termes de décision et de choix : toute forme de décision collective, tout suffrage peuvent être considérés comme l'agrégation de volontés et d'intérêts particuliers. Le suffrage, qui est ainsi fondamental, implique que le vote est un pari. Se pose alors une question : comment éviter au parieur les inconséquences possibles de son choix ? Ce que les sociologues nomment l'*effet Condorcet* montre que l'agrégation de jugements raisonnables donne naissance à des jugements déraisonnables et que la notion de volonté générale n'est pas si claire qu'on le croit à première vue. Le traitement mathématique de cette difficulté débouche sur deux catégories hétérogènes de problèmes : la définition d'une mesure d'utilité par une mesure de probabilité, et la construction d'un modèle d'homme rationnel dont on retrouverait le comportement dans les comportements spontanés.

En outre, chacun sait aujourd'hui que mathématiser n'est pas nécessairement quantifier, ce qui signifie ici que la capacité des mathématiques à servir d'instrument aux sciences humaines n'est pas toute concentrée dans la théorie des probabilités, avec ou sans l'auxiliaire des statistiques.

L'analyse des structures algébriques a eu une fortune qui a largement dépassé le domaine de l'algèbre; de Cl. Lévi-Strauss à Piaget, en passant par une multitude de travaux de linguistique et de sociologie, c'est le langage commun d'une mathématique de la structure qui domine, au point qu'on n'est pas loin d'espérer que les sciences de l'homme deviennent une branche des mathématiques appliquées. L'algèbre des structures fournit pour le moins des *modèles,* c'est-à-dire des représentations schématiques permettant l'étude d'un ensemble de questions, même si, comme c'est le plus souvent le cas, elles n'entretiennent pas de lien de signification naturel avec ces questions. La meilleure illustration de ce procédé est certainement fournie par l'étude des systèmes de parenté dans les sociétés archaïques : on se reportera, bien entendu, à la fameuse contribution, dans les *Structures élémentaires de la parenté* de Cl. Lévi-Strauss, d'André Weil, qui montre comment la théorie des groupes de substitution facilite la classification des lois du mariage dans la société Murngin (*cf.* chapitre XIV). L'analyse structurale des relations de parenté est fructueuse, précisément parce que la relation de parenté est éminemment sociale, mettant en jeu non pas deux ou trois individus isolés mais tout un ensemble plus ou moins grand d'individus, et à travers lui tout le groupe, c'est-à-dire toute l'organisation, toute la structure sociale.

La maternité, par exemple, est une relation non seulement d'une mère à ses enfants mais aussi de cette femme à tous les membres du groupe, pour qui elle est épouse, sœur, cousine, etc. Elle définit un ensemble de droits et de devoirs, et, comme le remarque si justement Lévi-Strauss, son absence ne définit pas rien, mais elle définit l'hostilité. Le mariage n'est pas « un processus discontinu, qui tire de lui-même, dans chaque cas individuel, ses propres limites et ses possibilités »; les règles qui le régissent expriment la façon dont un groupe donné organise l'échange et la circulation des femmes entre les différents segments de la société.

Ce qu'il y a de plus social dans la société est donc susceptible d'être mathématisé. Non que l'on trouve dans la société l'équivalent de véritables quantités théoriques : dans le meilleur des cas, en économie par exemple, la quantification statistique n'atteint que des mélanges; mais elle permet, néanmoins, de poser des questions dont le langage ordinaire ne permet pas une formulation précise. Par ailleurs, le recours de plus en plus courant aux modèles permet une exploration qui tient lieu, pour ainsi dire, d'expérimentation.

Mathématique et technologie

Il faut rappeler que cet accès progressif des phénomènes sociaux au traitement mathématique s'est fait dans une société qui se transformait parallèlement jusqu'à se muer en support d'une pensée essentiellement technologique, expression ultime de l'exigence de rationalité. Le terme *technologica* a d'abord désigné une doctrine de la division des disciplines, de leur classification; des acceptions proches du sens moderne apparaissent à la fin du XVII[e] siècle; par exemple, chez C. Wolf, il désigne déjà une science des métiers et de leurs produits.

Le premier représentant moderne de cette exigence, Descartes, concevait à la fois le monde comme un monde mathématique (qui ne contient qu'étendue et mouvement) et la science comme le moyen de s'en rendre maître. Désormais, l'expérience ne sera plus, comme dans l'Antiquité, aux antipodes de la spéculation; ayant elle-même changé de nature, son statut s'est modifié : conduite à l'aide d'instruments qui sont, selon une formule connue, des « théories matérialisées », elle s'est de plus en plus pénétrée de théorie. Cette tendance s'est accentuée à partir de la Révolution française : la perception humaine a définitivement troqué les sens pour les instruments. Dès 1794, Lakanal proclamait que sans mathématiques l'architecture civile et militaire n'a plus de règle et les sciences de l'artillerie et des fortifications plus de fondement. Et Monge de préciser qu'il faut orienter l'éducation nationale vers la connaissance des objets qui exigent de l'exactitude, vers les théories qui permettent des

applications précises. C'est le développement d'un nouveau modèle du savoir (les sciences appliquées), et l'exigence de nouvelles écoles (École centrale des travaux publics [1797]), où s'enseignent les mathématiques et la physique en vue de la diversité de leurs applications. Cournot remarquait que l'homme a fini par préférer à la machine naturelle, le cheval avec les images poétiques qu'il éveille, le cheval-vapeur de l'industrie moderne, qu'il a pu construire sur un plan plus simple, dont il peut mieux régler le service et contrôler la dépense. Dans ce but, il s'est fié de plus en plus aux mathématiques.

Cette technologie s'est étendue aujourd'hui sans cesser de nourrir un rapport profond aux sciences de la précision, c'est-à-dire plus ou moins directement aux mathématiques. Elle est l'application systématique des sciences, et fonde la possibilité de concrétiser le savoir sur une stratégie de la division et de la spécialisation extrêmes des tâches, caractéristiques qui sont, à leur tour, marquées par la durée des processus de fabrication exigeant une multitude de compétences généralement séparées. D'où le rôle, dans la civilisation technologique, de la *recherche* et de la *planification*. On n'en est plus à savoir calculer un certain nombre de forces physiques de l'Univers, mais à vouloir décupler, et défier en même temps ces forces; la concurrence internationale contraint à cet effort gigantesque si l'on prétend maîtriser l'espace, évaluer les matériaux disponibles, contrôler le savoir qui permet de transformer toutes ces données en autant de ressources.

Non seulement les mathématiques se sont imposées dans plusieurs disciplines longtemps réputées réfractaires, non seulement elles constituent un élément de base dans la société technologique industrielle d'aujourd'hui, mais encore, elles prétendent à une place de choix *dans la culture* actuelle. On est loin de l'époque où A. Comte se plaignait de la domination des Académies par les géomètres. On est loin de la mise en garde de Cournot contre la confusion du règne des chiffres, des nombres et des mesures avec celui des mathématiques. Il est désormais entendu que toute recherche sérieuse recourt d'une façon ou d'une autre aux mathématiques. La vieille bataille entre le latin (et le grec) et les mathématiques dans la formation des élèves n'a plus de raison d'être, car elle a été définitivement gagnée. Les mathématiques apparaissent comme le lieu où l'on peut disposer de toutes les formes possibles de discours, des discours fondamentaux qui sont à la racine de la mécanique, de la physique, de la biologie (il faut un baccalauréat de la série C pour des études de médecine), de la psychologie, de la sociologie, de la linguistique, etc. On apprend ainsi à raisonner sur tout en toute rigueur, à se méfier des idées sans expression immédiatement claire, à couler sa pensée dans le moule du raisonnement mathématique. Et surtout, celui qui est formé par l'activité mathématique laisse plus spontanément les généralités aux spécialistes de la politique pour accepter la sécurité intellectuelle que procure une discipline reconnue. Or, cette attitude est plus rentable pour une société technologique, qui divise et décompose tout problème en segments multiples dont chacun constitue le domaine d'intervention d'un expert. Les contraintes de l'industrie et, en dernier ressort, le marché du travail renforcent la préférence pour la culture mathématique. Mais cette préférence est une préférence dictée; c'est le résultat de la contrainte sociale, d'un fait social au sens plein du terme, qui impose, comme eût dit E. Durkheim, « des manières d'agir, de penser et de sentir extérieures à l'individu, et qui sont douées d'un pouvoir de coercition en vertu duquel elles s'imposent à lui », sans avoir ni la nécessité des faits logiques ni le caractère naturel de ce qui s'enracine dans l'organique, ni le caractère personnel propre aux représentations psychologiques.

L'existence de ce fait social légitime l'effort de faire une véritable sociologie des mathématiques, que nous ne prétendons pas développer entièrement mais en vue de laquelle nous voulons faire les remarques, à caractère programmatique, qui suivent.

Réflexions pour une sociologie des mathématiques

Une sociologie des mathématiques ne saurait être considérée comme un appendice à la sociologie de la connaissance générale, qui est l'étude des formes de connaissance propres à chaque société ou à chaque groupe, et qui repose sur une notion trop générale pour être opérante dans une enquête restreinte. Mais on ne saurait non plus l'assimiler à une sociologie de la science, d'une part parce que les rudiments déjà disponibles de cette sociologie ne considèrent pas les mathématiques comme un secteur particulier digne d'une recherche de ce genre, d'autre part parce que sous la rubrique « sociologie des sciences », qui recouvre un vaste domaine, depuis une sociologie de la sociologie elle-même à une sociologie de la communauté savante, on ne considère généralement que les sciences où l'impact social est facile à déceler, en raison des services multiples qu'elles rendent directement à la société industrielle. Or les mathématiques pures restent un peu à part, du fait qu'elles ont une autonomie relative par rapport à la société, et nous allons préciser en quel sens pour éviter tout malentendu.

Rapports des mathématiques aux formes sociales

La préhistoire et l'Antiquité

En mathématiques, l'investigation peut être comprise sans qu'il soit nécessaire de recourir à un point de départ empirique. Ni l'observation des faits, qu'ils soient fortuits ou produits, ni l'émission d'une idée ou d'une hypothèse ne constituent des considérations propres à éclairer l'activité de recherche du mathématicien. L'opinion qui place l'empirisme, c'est-à-dire l'observation ou l'expérience fortuite, à l'origine de toutes les sciences, se rencontre encore aujourd'hui, sous une forme raffinée certes, et s'exprimant par des moyens on ne peut moins empiriques il est vrai, comme le montrent les travaux de certaines tendances de l'école analytique des dernières décennies. L'empirisme le plus intégral ne peut, cependant, méconnaître le statut du formel : aussi la mathématique, ne pouvant procéder de l'expérience, se voit-elle reléguée dans une sphère de pur formalisme, après quoi se pose évidemment la question de la fécondité d'une science purement formelle !

Une sociologie des mathématiques ne peut avoir pour but de fonder les mathématiques ou d'en donner une explication ultime. Il ne peut être question pour elle d'assumer ce projet philosophique au moment où les mathématiques ne s'embarrassent plus d'aucune vue systématique ni d'aucune philosophie ! Par ailleurs, une fondation sociale de cette science pourrait tout simplement balancer dans la contradiction ou le paradoxe une activité dont le premier but est de s'en garder. En mathématiques, le sociologisme est aussi peu à sa place que le psychologisme; il n'y a pas de genèse sociale, pas plus qu'il n'y a de genèse psychologique de la mathématique.

L'histoire des mathématiques montre qu'à tel moment donné, tous les résultats ne connaissent pas une égale fortune : certains restent en friche, tandis que d'autres, favorisés par une vive curiosité ou un grand intérêt social, sont aussitôt exploités. Par ses encouragements, la société peut ainsi jouer un rôle moteur, susciter le développement ou favoriser le plein épanouissement d'une science dont la facture interne ne reflète pas immédiatement les besoins et les passions sociales.

Examinons les choses d'un peu plus près. L'archéologie nous apprend qu'il n'y a probablement pas eu, avant l'Age du bronze, de pratique mathématique diversifiée, bien que l'on ait connu dès la fin du Néolithique des rudiments de calendrier et certaines figures géométriques. Remarque dont il faut souligner le caractère approximatif, en rappelant l'absence d'une histoire unique pour toutes les aires culturelles, et le fait que ce qui est découvert ici vers 3500 avant J.-C. n'apparaît ailleurs que vers 500. L'Age du bronze produit donc un calendrier plus précis, des rudiments d'arithmétique, de géométrie, d'astronomie, mais c'est aussi l'âge où apparaissent la propriété privée, la société de classes, et surtout une classe distincte de celle des paysans et des artisans, la classe des scribes. L'époque suivante connaît l'alphabet, mais voit également apparaître la classe des commerçants et des formes politiques plus élaborées; autrement dit, on assiste au développement du secteur de l'échange et de celui de l'administration; l'alphabet consacre l'écriture, et l'écriture crée, dit-on, l'histoire.

Considérons le cas de l'Égypte : sa population apparaît dès le Néolithique; ses outils sont en pierre polie, elle connaît la poterie, la vannerie, le tissage, le travail du cuir, du bois et de l'os; elle a domestiqué le bœuf, le mouton,

la chèvre, et cultivé le blé et l'orge, d'où une première organisation sociale. Après le Néolithique, les céramiques témoignent d'un sens géométrique, telle cette céramique qui présente sur fond rouge poli un décor géométrique peint en blanc.

Ailleurs, en Abyssinie et en Mésopotamie, on remarque une évolution similaire, mais les spécialistes refusent de parler d'influence. Cette évolution, en tout cas, trahit un perfectionnement des techniques qui a transformé l'artisan en artiste : une première division du travail lui a sans doute permis de développer, en se consacrant exclusivement à son travail, un goût esthétique, un sens du parfait qui prévaut sur le sens de l'utile et approche l'idée d'objet d'art. A l'époque où existe une division technique du travail qui est, en raison de l'importance religieuse du calendrier, une division sociale du travail, apparaissent également des notions fragmentaires d'arithmétique et de géométrie. Entre les mains d'une caste, celle des scribes, l'art de calculer, d'abord lié à la métrologie habituelle (mesure des longueurs, des volumes, des poids, etc.), aux nécessités vitales de l'agriculture (prévision des crues du Nil, restitution des frontières des terrains détruits par l'inondation), enfin aux exigences de la vie administrative, tel l'établissement juste d'une assiette de l'impôt, s'est développé sans s'arracher aux préoccupations utilitaires et empiriques qui semblent l'avoir suscité. Les scribes, déjà au fait des techniques de rédaction, se chargent des techniques de calcul et de mesure, produit d'un travail appliqué mais autonome, déjà spécialisé, qu'il est difficile de remiser dans l'expérience au sens de l'empirisme traditionnel.

Le fait n'a pas échappé à Aristote qui, nonobstant le style archaïque, a su l'apercevoir, mieux que tous ceux qui, pour défendre l'originalité absolue du miracle grec, n'ont pas voulu accorder à l'Égypte ce que le grand philosophe grec lui reconnaissait volontiers : qu'elle a été « le berceau des arts mathématiques, car on y laissait de grands loisirs à la caste sacerdotale ». Les arts mathématiques sont donc nés dans les contrées « où régnait le loisir », c'est-à-dire où certains travailleurs pouvaient se consacrer à une activité qui ne répond pas immédiatement aux nécessités de la vie. La caste des scribes n'était directement soumise qu'à l'administration et à la religion !

Les techniques de calcul et de mesure ont donc un certain *caractère officiel*, que l'on retrouve, plus accusé encore, dans les remarquables réalisations chinoises, caractérisées, pour ce qui concerne les mathématiques, par la prédominance de la pensée algébrique. En Chine, plus qu'ailleurs, le développement d'une bureaucratie au service de l'appareil d'État, pour organiser l'ensemble de la production et diriger la main-d'œuvre dans les travaux agricoles et de construction, une bureaucratie bénéficiant d'un « charisme impérial » à la mesure de son efficacité dans l'organisation sociale, semble avoir joué un rôle décisif dans la formation de la pensée mathématique. La société chinoise ancienne n'est pas fondamentalement esclavagiste, et, à la différence de la société égyptienne, prompte à utiliser massivement la force humaine, utilise plutôt des moyens techniques, telle la voile pour la propulsion des bateaux. Mais l'autorité centrale ne s'exerce qu'avec le concours des savants qui se trouvent aux commandes et cherchent à tirer parti du cours naturel des choses en intervenant aussi peu que possible dans les affaires de la société. Cette conception *non interventionniste* dans le champ de l'activité humaine aurait dû favoriser le *développement des sciences de la nature*. Mais, et c'est là la limite de la science chinoise, la pérennité de la féodalité bureaucratique, qui se conserve sans changement, peut avoir constitué un obstacle à l'association des mathématiques avec l'observation empirique. La science chinoise n'accomplit pas le saut qualitatif qui s'accomplira dans la science occidentale, où l'expérimentation, exigeant l'intervention active du savant, scellera à partir de Galilée une association décisive avec les mathématiques.

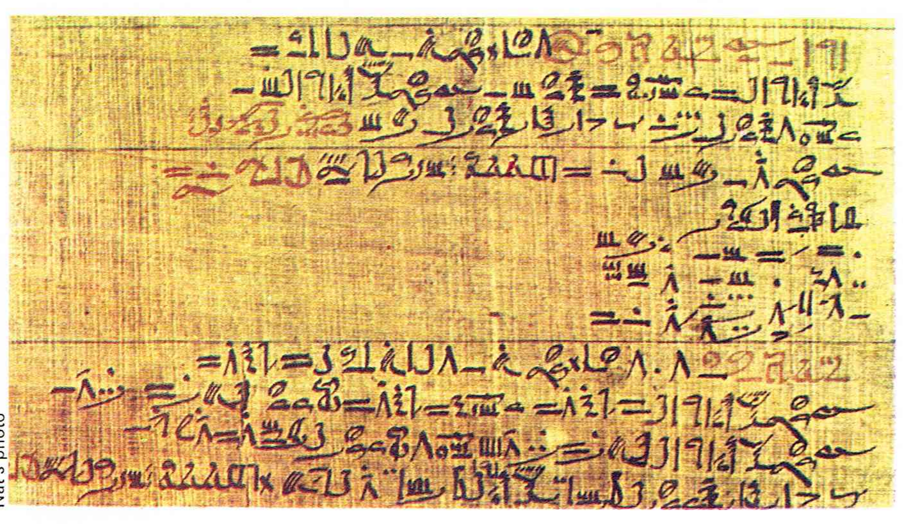
On ne peut évidemment pas dégager ici toutes les corrélations possibles, pour chacune des grandes formes de vie collective, entre le développement d'une certaine mathématique et la présence de certains facteurs dans l'organisation sociale. Mais le peu que nous en avons dit montre la faiblesse des facteurs anthropo-physiques, raciaux ou intellectuels devant les *facteurs sociaux*. Malheureusement l'étude sociologique ne peut être menée qu'avec une extrême prudence et un sens très vif des nuances, et ses résultats auront difficilement la forme d'énoncés catégoriques, car rien ne serait plus hasardeux, comme nous l'avons déjà fait remarquer, qu'une mise en parallèle de la progression en mathématiques et des formes sociales qui impliquerait un découpage aussi dogmatique qu'intellectuellement fragile.

Que les sciences mathématiques les plus abstraites se soient épanouies dans la Grèce antique, c'est sans doute parce que le loisir des uns a permis de dégager l'activité mathématicienne des pratiques directement commandées par la vie religieuse, commerciale et technique. Ce qui explique, notamment, l'attribution aux mathématiques de qualités esthétiques indépendantes, d'un genre de beauté particulier, effet des rapports d'ordre, de proportion et d'harmonie qu'y découvre un Pythagore, capable de voter une hécatombe le jour où il aperçoit, sans le secours d'aucun arpenteur, la relation qui lie les trois côtés d'un triangle rectangle. Ce caractère répond à des facteurs d'ordre sociologique, dont une fiction pourra indirectement montrer l'importance : imaginons Monge, par exemple, expliquant ses procédés de géométrie descriptive à quelque architecte des temples d'Agrigente ou de Sybaris, contemporain de Pythagore, qui n'aurait vu dans cette géométrie, si décisive pour le XIXe siècle européen, qu'un artifice ingénieux, trop particulier à l'homme, à son industrie, aux instruments matériels dont il dispose.

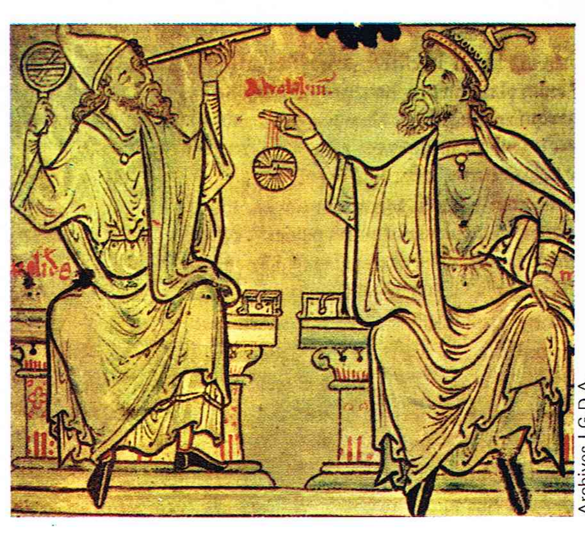
Le tournant du XVIIe siècle

On voit, par contraste, le caractère éminemment social des mathématiques d'aujourd'hui, héritières de celles qu'un certain état de l'Europe à la fin du Moyen Age a rendues possibles : une liaison spécifique entre les mathématiques et l'expérience, plus précisément une mathématisation des hypothèses de l'expérience qui a mis celle-ci sous le contrôle de celles-là, a permis de lier le sort des mathématiques à celui de la physique et de les transformer

▼ A gauche, détail d'un papyrus égyptien indiquant la méthode de calcul du volume d'un réservoir cylindrique à blé, connaissant le diamètre et la hauteur. A droite, dans les* Éléments *d'Euclide (représenté en une double figuration dans ce manuscrit médiéval), la géométrie prend un aspect d'une beauté toute particulière et d'un enchaînement logique.



Nat's photo



Archives I.G.D.A.

en magasin d'outils anticipatoires exigés par les différents laboratoires techniques. Si l'on veut éprouver la teneur de cette thèse, le concept de « formation du capitalisme moderne » se présentera naturellement à l'examen, et dans l'acception étendue que l'expression a invinciblement acquise depuis le début de ce siècle, à savoir celle d'une rationalité moderne, dont les causes multiples et variables ont été diversement commentées. Ce qui est certain, c'est que les mathématiques pures d'aujourd'hui résultent d'un effort de rigueur et de précision soutenu tout au long du XIXe siècle, centré sur l'analyse, qui est née de l'inspiration physique et mécanique présente depuis la coupure galiléenne. Engels avait raison de dire qu' « en introduisant les grandeurs variables et en étendant leur variabilité jusqu'à l'infiniment petit et à l'infiniment grand, les mathématiques aux mœurs si austères ont commis le péché », si l'on entend par là qu'elles se sont ouvertes à une histoire.

Après quoi la validité de leurs démonstrations sera à revoir; un nouveau système sera nécessaire pour donner aux nouvelles notions un statut logique indépendant du sens qu'elles ont en naissant. Mais du point de vue d'une sociologie historique, la période caractéristique des mathématiques est bien ce siècle qui, en étant celui de Galilée et de Descartes, fut celui de la navigation outre-mer, du négoce entreprenant, des sciences physiques, des premières statistiques importantes, celui du début de la mécanisation de l'industrie et des principaux phénomènes spécifiques de la civilisation industrielle.

Le scientisme actuel

Aussi est-ce avec raison que mathématiciens et non-mathématiciens à la fois insistent sur les succès spectaculaires de la méthode expérimentale pour expliquer l'impact des mathématiques sur la vie sociale et quotidienne. Si bien que quiconque veut être de la société moderne, ou moderne tout court, qu'il soit homme d'affaires, politicien, ou linguiste, doit disposer d'un bagage mathématique suffisant pour être à l'aise dans la langue des chiffres. Ce phénomène, né dans une société donnée, est entretenu par la publicité faite aux mathématiques par une idéologie propre : le scientisme, nouvelle religion servie par les prêtres respectés que sont les scientifiques, les technocrates, les experts, unanimes à réagir à « toute attaque contre cette religion, ou l'un de ses dogmes, ou l'un de ses produits, avec toute la violence émotionnelle d'une élite régnante aux privilèges menacés ».

Cette élite, remarque encore A. Grothendieck que nous venons de citer, s'identifie intimement aux pouvoirs en place qui s'appuient fortement, en retour, sur ses compétences technologiques et technocratiques. C'est elle qui porte et propage l'ensemble des mythes qui forment le credo scientiste : que seule la connaissance scientifique est véritable, réelle, objective, universellement valide; que tout ce qui peut être exprimé sous forme quantifiée et produit dans des conditions expérimentales déterminées, est objet de connaissance scientifique; que toute la réalité doit se laisser exprimer par des modèles mécaniques ou formels, analytiques en tout cas; que la connaissance doit être pulvérisée en districts spécialisés où seuls interviennent et jugent ceux qui en sont les spécialistes et les experts; que science et technologie sont capables de résoudre les problèmes de l'homme; enfin, que les experts seuls sont en mesure et méritent d'appartenir aux instances de décision.

Ainsi la mathématique n'est pas seulement descendue dans l'arène sociale, elle remplit directement une fonction sociale. Et d'abord par le biais de la technologie, ce qui se traduit dans le rôle de l'ingénieur et du technocrate dans notre société. Personnages issus probablement des bâtisseurs de cathédrales, qui semblent bien avoir été les premiers à réunir connaissance scientifique et connaissance technique, qui ont, les premiers, permis aux patrons de ne plus conduire personnellement leurs travaux et qui ont scandalisé, dès cette première apparition, Nicolas de Biard écrivant en plein XIIIe siècle : « Dans les grands édifices, il y a un maître principal qui les ordonne seulement par la parole, mais n'y met que rarement ou jamais la main, et cependant il reçoit des salaires plus considérables que les autres... Les maîtres des maçons ayant en main la baguette et les gants disent aux autres *Par ci me taille,* mais, eux, ils ne travaillent et cependant ils reçoivent une plus grande récompense. » Ou plutôt ils travaillent autrement et constituent les premiers spécialistes auxquels on fait appel, parfois de fort loin.

Cette situation d'imbrication mutuelle entre des exigences nées dans des domaines aussi hétérogènes que celui de la mathématique et celui de la vie sociale est propre à expliquer la nature psycho-sociale des *paradigmes* que Thomas Kuhn présente dans son essai *Structure des révolutions scientifiques :* à tel moment tel secteur scientifique s'impose comme sommet et guide des recherches poursuivies sous l'effet des contraintes qu'impose la société. Ce qui n'est possible que parce que le scientifique, fût-il pur mathématicien, est d'emblée, à sa naissance, mêlé, voué au service social, ainsi que l'exprime le souhait des classiques, de Descartes entre autres, lorsqu'il déclare : « Pour ce qui est des expériences qui peuvent y servir, un homme seul ne saurait y suffire à les faire toutes; mais il ne saurait aussi employer utilement d'autres mains que les siennes, sinon celles des artisans, ou telles gens qu'il pourrait payer, et à qui l'espérance du gain, qui est un moyen très efficace, ferait faire exactement toutes les choses qu'il leur prescrirait » (*Discours de la méthode,* VIe Partie).

L'impact de l'État

Or, le lien de la mathématique à la société est aujourd'hui plus profond que jamais, le savant dépendant lui-même d'une instance de décision qui lui commande de travailler dans le sens d'une mathématique significative, c'est-à-dire utile à la société industrielle. C'est dans ce sens assurément que se fait l'évolution, comme le montrent les informations les plus récentes :

Nous apprenons qu'à l'occasion du 20e anniversaire de l'Académie, L. Brejnev a déclaré aux savants russes : « Nous n'avons pas l'intention de vous dicter les détails des thèmes scientifiques, les voies et les méthodes de la recherche. C'est l'affaire des savants eux-mêmes. Mais quant aux orientations essentielles du développement de la science, quant aux tâches principales rendues nécessaires par les réalités, nous les déterminerons ensemble. » Autrement dit, les experts doivent réaliser des programmes dont ils ne décident qu'à titre, au plus, de conseillers.

La situation n'est guère différente ailleurs. En France, le gouvernement a décidé une série de réformes de structure visant la coordination technico-scientifique; désormais « les laboratoires élaboreront des programmes, mais c'est le gouvernement qui disposera, en s'aidant de comités consultatifs et de la Délégation générale à la recherche scientifique et technique ».

L'inféodation à l'État-Patron et aux sociétés industrielles « intègre » la vie scientifique à la vie économique. Elle consacre une dépendance étroite des sciences, et en premier lieu des mathématiques, envers la société. C'est ainsi que les mathématiques dites « modernes » ne sont pas dénuées d'attaches idéologiques spécifiques; René Thom a souligné, par exemple, que l'esprit bourbakiste, essentiellement algébriste, tend à faire passer dans l'enseignement les structures algébriques qui s'étaient révélées utiles dans les mathématiques récentes, en vue d'une formation mathématique accessible à des hommes ordinaires, qu'on informe minutieusement, et qu'on mène, au moyen d'un enseignement approprié, à faire en sorte que leurs connaissances se combinent avec celles de spécialistes d'autres branches, également ordinaires, pour produire l'immense littérature fabriquée dans les départements de mathématiques, équivalents sobres des laboratoires de physique.

Les mathématiques n'ont donc pas servi seulement à rationaliser l'étude des secteurs sociaux. Par le biais de la technologie et des exigences de la société industrielle avancée, elles se trouvent au principe même de la société. Au point que celle-ci, pour en produire, se dispense des génies, et se suffit des hommes ordinaires, convenablement formés ! Qu'est-ce à dire sinon que la mathématique n'est pas le produit d'idiosyncrasies individuelles — et l'a-t-elle jamais été ? Programmes, groupes de recherches, autant d'expressions qui nous rappellent que la mathématique est tombée dans la perspective sociale. Les programmes impliquent des paradigmes qui définissent les problèmes à résoudre, et ces paradigmes définissent comme une tradition horizontale, où l'on n'imite point ses pères, mais les directeurs de programmes et de recherches, dépositaires des ultimes suggestions réalisables parmi toutes les suggestions scientifiquement possibles.



Giraudon

▲ *Le mathématicien Johann Neudörffer et son fils (portrait dû au peintre flamand Nicolaas Neufchâtel).*

MATHÉMATIQUES ET PÉDAGOGIE

L'éducation mathématique se place aujourd'hui au premier rang des fins que se propose l'enseignement. Seul l'enseignement de certaines langues vivantes pourrait concourir avec celui des mathématiques, mais comme ce dernier est plus délicat à concevoir, plus difficile à dispenser, il tient, semble-t-il, la première place dans les préoccupations de la pédagogie moderne. On pourrait même dire qu'en France, la pensée pédagogique a connu un certain renouvellement grâce aux problèmes posés par l'enseignement des mathématiques, tandis que, dans d'autres pays, la pédagogie possède un statut institutionnel régulier et permanent, comme c'est par exemple le cas dans les universités allemandes.

La relance de cette réflexion est profondément enracinée dans les exigences de la société actuelle; nous avons souligné dans le chapitre *Mathématiques et société* que la connaissance de notions élémentaires de mathématiques est exigée de tous, ou si l'on veut, de tous les cadres de la société. L'élite, de plus en plus, est appelée à apprécier le point de vue numérique et à être préparée de manière à savoir prendre des décisions techniquement, c'est-à-dire quantitativement, fondées. En particulier, on exige des chercheurs et des ingénieurs des connaissances mathématiques plus solides, en raison des nouvelles applications de la mathématique à divers secteurs de l'activité économique.

Mais quels sont les faits qui sont impliqués dans cette réflexion pédagogique?

— Il faut remarquer, tout d'abord, que nous vivons encore dans la postérité d'une découverte qui constitue l'événement de la pédagogie moderne : la découverte de l'enfant. La découverte d'un être qui n'impose pas seulement une réflexion sur le contenu des programmes mathématiques, mais aussi une élaboration des techniques de transmission du savoir; l'enseignement des mathématiques pose, au niveau élémentaire, des problèmes psychologiques qu'il faudrait formuler correctement et, si possible, résoudre.

— Sur ces problèmes généraux, se greffe un problème interne aux mathématiques qui ont leur querelle des anciens et des modernes. Il ne s'agit pas seulement d'enseigner les mathématiques dans le dessein de faciliter l'insertion sociale, mais aussi de savoir quelle mathématique enseigner, dans quel style, avec quelle conception de la rigueur. Et faut-il préparer dès le début l'enfant à la matière qu'il devra connaître s'il se spécialise plus tard en mathématiques, ou faut-il seulement l'exercer à penser mathématiquement, à assimiler moins les notions qu'un art de penser?

— Enfin, l'évolution de l'École, qui s'impose aujourd'hui à un plus grand nombre d'enfants qu'auparavant, a engendré l'exigence que tous soient également pris en charge par l'enseignement et que personne ne soit — en principe — « laissé pour compte ».

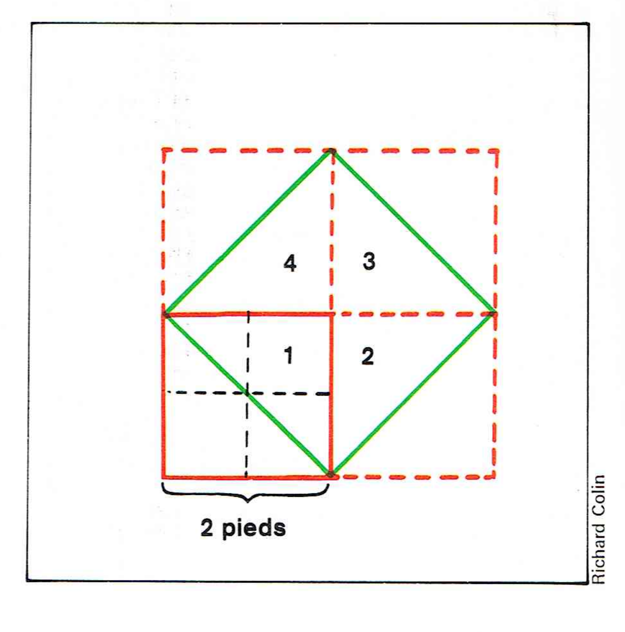
Il va de soi que ces problèmes ont entre eux certains rapports : la formation des professeurs dépend étroitement de la question de savoir quelles mathématiques enseigner; le problème de savoir comment introduire les mathématiques a une relation directe avec celui que posent les sujets auxquels s'adresse l'enseignement; et surtout, on ne trouvera de réponse claire à aucune question si on ne résout celle des buts à atteindre par cet enseignement.

La naissance du concept d'enfant

Le problème purement pédagogique que posent les mathématiques se situe au niveau de leur enseignement aux plus jeunes enfants. La première difficulté est de savoir comment articuler, sans une trop grande distorsion, un entendement mathématique (c'est-à-dire l'aptitude à exercer une activité intellectuelle qui prépare à une pratique ultérieure des mathématiques) sur les possibilités qui sont propres à l'enfant.

Poser cette question suppose qu'on s'est heurté à des échecs fréquents, et à travers eux, à la réalité spécifique de l'enfant, que la pédagogie mathématique, longtemps réduite à une simple didactique, c'est-à-dire à une technique détachée de toute contexte psychologique, historique et social, ne peut plus ignorer aujourd'hui.

Traditionnellement, on oscillait entre deux voies : le psittacisme qui fait appel à la mémoire répétitive de l'enfant pour fixer les règles de calcul, dont l'origine remonte au Moyen Age et à sa prédilection pour la versification mnémotechnique, et dont l'apprentissage des tables de multiplication a retenu la leçon d'efficacité; et la maïeutique, immortalisée par le dialogue entre Socrate et Ménon. Celle-ci est à l'origine d'une conception, non disparue aujourd'hui, d'une mathématique éternelle et indépendante. La maïeutique tient tout entière dans le mot fameux : « C'est toi qui le diras »; mais on remarquera qu'il s'agit, en fait, d'une méthode fondée sur le déploiement d'un doute, c'est-à-dire sur le rejet systématique des « préjugés » liés à la langue naturelle : Ménon, qui a pourtant mille fois parlé de la vertu, se trouve, après un premier entretien avec Socrate, totalement désorienté et ne sait plus que penser. C'est le moment que choisit



Richard Colin

▶ *Le problème du carré soumis à l'esclave de Ménon par Socrate.*

Socrate pour affirmer qu'apprendre, c'est se ressouvenir, et le montrer pour ainsi dire expérimentalement, en faisant appel à un des esclaves de la nombreuse suite de Ménon. Celui-ci, de qui on n'exige qu'une chose, entendre le grec, se montre capable de trouver comment obtenir un carré dont la surface est double de celle d'un carré donné de deux pieds. Socrate n'apporte pas du dehors la science à son nouveau disciple, mais l'aide à se représenter visuellement la figure à construire qui sera sans doute un carré de côté plus grand que celui du carré donné mais plus petit que celui du carré de quatre pieds de côté, auquel a d'abord songé l'esclave. Peut-être, pense alors l'esclave, la solution consiste-t-elle à prendre un côté de trois pieds, mais il est vite détrompé. Devant la difficulté, force est de revenir à la première hypothèse et de regarder comment se présente le carré de quatre pieds de côté; sa surface vaut quatre fois celle du carré donné au départ : le dessin montre les quatre carrés égaux au premier carré; mais si on divise chacun de ces quatre carrés en deux triangles égaux, en traçant les diagonales convenables, on voit que les diagonales forment à leur tour un carré qui répond à la question. L'esprit de l'esclave a cheminé progressivement vers cette vérité sans que Socrate ait auparavant exigé de lui autre chose qu'une communauté de langue permettant le dialogue.

Si nous avons rappelé cette méthode, c'est qu'elle contient tous les problèmes posés par l'enseignement des mathématiques. Que l'usage du dessin, de la figure, ou, si l'on veut, de l'intuition, n'explicite pas ses règles est un problème de didactique que nous examinerons plus loin; mais la façon dont intervient la notion de diagonale pose le problème de l'introduction d'une notion accessoire : la façon dont on le résout implique une définition de la science, laquelle apparaît effectivement plus loin dans le *Ménon*, dans le sort qui est fait à l'idée d'*hypothèse :* quand on demande au géomètre à propos d'une surface, par exemple, si tel triangle peut s'inscrire dans tel cercle, il répondra : « Je ne sais pas encore si cette surface s'y prête; mais je crois à propos, pour le déterminer, de raisonner par hypothèse de la manière suivante : si cette surface est telle que le parallélogramme de même surface appliqué à une droite donnée est défaillant de telle surface, le résultat sera ceci; sinon il sera cela. »

La mise en forme de la conception sous-jacente à cette pédagogie a donné les *Éléments d'Euclide,* chef-d'œuvre de logique sans pédagogie, fondé sur l'idée que pour les mathématiques il n'y a point de voie royale, ni de voie enfantine.

Rousseau fut sans doute le premier à en dénoncer explicitement et systématiquement le caractère abstrait. Non seulement il déclare, dans les *Confessions,* ne point goûter personnellement la méthode d'Euclide qui cherche plutôt la chaîne des démonstrations que la liaison des idées, non plus que celle de l'algèbre « abstraite », mais encore, dans l'*Émile,* il dénonce la tentation de suggérer ou de dicter des démonstrations sous le couvert d'un apprentissage de la découverte. Nul n'a trouvé avant lui de formule aussi nette pour souligner la nécessité d'élaborer une pédagogie réelle : « Nous ne sentons pas, écrit-il, que leur méthode [celle des enfants] n'est pas la nôtre, et que ce qui devient pour nous l'art de raisonner ne doit être pour eux que l'art de voir. Au lieu de leur donner notre méthode, nous ferions mieux de prendre la leur. » Il ne s'agit de rien moins que de l'effacement de l'éducateur devant l'enfant et de la mise en question du fait que l'éducation est généralement moins une formation du jugement que la stratégie de reproduction d'une mentalité adulte donnée.

Un véritable *concept de l'enfant* est créé; l'éducation y sera subordonnée dans la mesure où l'enfant n'a pas à apprendre quelque chose mais à apprendre à apprendre. D'où la critique acerbe de « la manie enseignante et pédantesque », confinée dans la transmission des sciences qui sont le moins utiles à l'enfant, alors que par une véritable providence les facultés ne se développent qu'avec les occasions d'être exercées, en sorte qu'elles ne sont jamais « ni superflues ni à charge dans le temps, ni tardives et inutiles au besoin ». Les facultés humaines sont donc, selon Rousseau, d'abord des facultés virtuelles, et le développement de la raison est à son tour lié à celui de la sociabilité. D'où la nécessité de respecter « l'enfance » qui n'est encore capable ni de jugement, ni de raison, ni de mémoire; et donc point d'initiation directe aux mathématiques. « En m'objectant qu'ils apprennent quelques éléments de géométrie, on croit bien prouver contre moi; et tout au contraire, c'est pour moi qu'on prouve : on montre que, loin de savoir raisonner d'eux-mêmes, ils ne savent pas même retenir les raisonnements d'autrui; car suivez ces petits géomètres dans leur méthode, vous voyez aussitôt qu'ils n'ont retenu que l'exacte impression de la figure et des termes de la démonstration. A la moindre objection nouvelle, ils n'y sont plus. Renversez la figure, ils n'y sont plus! Tout leur savoir est dans la sensation, rien n'a passé jusqu'à l'entendement » *(Émile, livre II)*.

Par là, Rousseau est certainement le premier pédagogue : il crée une préoccupation visant l'enfant en tant que tel, non sans rester lui-même prisonnier de l'idéologie de son siècle, celle de Condillac surtout, qui le conduit à comparer l'enfant à une « table rase », apte à recevoir des connaissances « contingentées », illustrées et adaptées aux besoins de chaque étape d'un développement dont toute l'histoire baignait encore dans le mystère. Mais, désormais, la nécessité est posée qu'éduquer l'enfant implique d'abord de le connaître. A vrai dire, ce principe ne fut mis en œuvre que dans les dernières décennies du XIXe siècle, où les nouvelles exigences politiques et démographiques ont conduit à ce qu'É. Claparède (1873-1940) considérait comme une révolution copernicienne : le fait de centrer l'école sur l'élève au lieu de mettre l'élève au service de l'école. En réalité, il s'agit de la première « institutionnalisation des droits de l'enfant » également proclamés par la Révolution (lois des 16 et 24 août 1790 et décrets de 1792 et 1793).

L'intérêt pour l'enfance dépasse alors de loin le plan conceptuel et idéologique auquel étaient restés Rousseau et les déclarations révolutionnaires; il se concrétise dans l'étude « scientifique » de l'enfant. W. T. Preyer ouvre la voie par une étude consacrée à l'esprit de l'enfant (1881), bientôt suivie (1891) aux États-Unis par l'initiative décisive de G. S. Hall qui fonde un périodique spécialisé, puis le « National Association for the Study of Children ». Des institutions analogues se créent un peu partout en Europe, et en Russie. En France, F. Buisson et A. Binet animent la « Société libre pour l'étude psychologique de l'enfant »; à Genève, l'« Institut Jean-Jacques Rousseau » reprend, dès 1912, le travail commencé dans le séminaire de Claparède créé en 1906.

Mais le renouvellement de la conception de l'enfance ne sera total qu'avec les révélations de la psychanalyse, qui voit en l'éducation la répression organisée des instincts, répression rarement réussie, et toujours « au profit d'un petit nombre d'hommes privilégiés dont il n'a pas été requis qu'ils réprimassent leurs instincts ». L'éducateur ne peut plus ignorer les difficultés d'une voie à chercher entre « le Scylla du laisser-faire et le Charybde de l'interdiction ».

Ainsi, on ne peut éduquer un enfant en dépit de son « affectivité »; d'autre part, toute « nourriture intellectuelle » n'est pas bonne indifféremment à tous les âges; il faut tenir compte des intérêts et des besoins de chaque période.

L'enseignement doit donc s'adapter; et l'enseignement des mathématiques ne peut faire exception. Mais l'adaptation est d'autant plus juste que l'on a su détecter les aptitudes ou les incapacités. C'est A. Binet qui consacre l'approche métrologique de l'écolier par l'évaluation de son intelligence et l'élaboration d'échelles, de calcul et de lecture notamment. Cela conduit à *l'échelle métrique de l'intelligence,* de l'inintelligence diront d'autres! qui a été reprise par R. Zazzo en 1949 et en 1966, et qui s'est imposée par l'intermédiaire des notions aujourd'hui consacrées d'*âge mental* et de *quotient intellectuel*.

Pour ce qui concerne les mathématiques, ce qui est au centre de la réflexion pédagogique, c'est le problème que constitue l'incompréhension des enfants; les mathématiques seraient pour certains élèves des activités « chargées d'angoisse » (voir Stella Baruk, *Échec et Maths,* Seuil, 1973), ce qui se comprend aisément quand on voit « dans quelles aventures les entraînent l'échec scolaire d'une manière générale, et l'échec en maths en particulier. De bilan en test et de test en dossier, de dossier scolaire en dossier psycho-scolaire, de dossier psycho-scolaire en consultation de psychologue, et de consultation en rééducation, le chemin est jalonné d'épreuves, de contre-épreuves, de diagnostics et de contre-diagnostics convergents ou contradictoires, qui ont cependant en commun

le superbe non-sens : expliquer l'échec par sa justification ». C'est dire la déconvenue de tous les recours de la pédagogie à la psychologie, qui vont généralement dans le sens de cette justification, et qui marquent, sur le plan historique, certains moments forts : les crises manifestées par l'échec de l'institution scolaire dans ses projets de transformation du système éducatif. D'abord, la faillite de « l'illusion pédagogique » de la démocratie scolaire a conduit à la création de l'arsenal conceptuel qui préside à l'usage idéologique de la notion de « débilité » ; la découverte, après l'augmentation du nombre des élèves exigée par l'avancement de la société industrielle, du rapport entre réussite scolaire et origine sociale a motivé les études sur la sociologie de l'éducation, d'abord aux États-Unis (Skelles et Fillmore dès avant la Seconde Guerre mondiale), puis en France (R. Zazzo et Dabout ont fait en 1954 une publication sur la progression scolaire et l'inégalité des enfants devant l'école). Ensuite, l'idée que toute pédagogie dépend, en dernier ressort, d'un choix idéologique et politique, qu'elle vise chaque fois, et non seulement dans des cas limites, à réduire le *désir* de l'enfant au *besoin*, en lui substituant des automatismes qu'elle considère comme autant de garde-fous, autrement dit, la rencontre dans leurs ultimes exigences des critiques d'origine marxiste et psychanalytique, a conduit souvent à un syncrétisme décevant, mais parfois aussi à une meilleure analyse des raisons de l'échec des conseillers d'orientation scolaire et professionnelle, institués après la dernière guerre, et de tout le système d'éducation qui produit l'échec scolaire. Au total, l'école apparaît comme un appareil idéologique au service de l'État, qui ne fait que reproduire les rapports sociaux existants. Comme on l'a dit, c'est dans les fonctions sociales qu'on trouvera les raisons de la structure de l'école.

L'échec mathématique ne peut échapper à la force d'analyses de ce type ; elles nous disent que l'échec de l'enfant est un mythe, surtout lorsqu'il s'agit de mathématiques, où l'on s'ingénie à faire croire que l'enfant peut en faire la découverte personnellement et par lui-même, quand le moindre concept a coûté des années de travail à un ou plusieurs auteurs ! Comment servir à un enfant le concentré de recherches millénaires ? Cette question, en tant qu'elle manifeste le souci de l'enfant, pose des problèmes de didactique mathématique que nous allons examiner maintenant.

▼ Cette mosaïque pompéienne du IIe siècle avant J.-C. représente Platon enseignant la géométrie, témoignage éloquent de l'importance attribuée à cette science par le monde grec.



Giraudon

La didactique mathématique

La maïeutique platonicienne

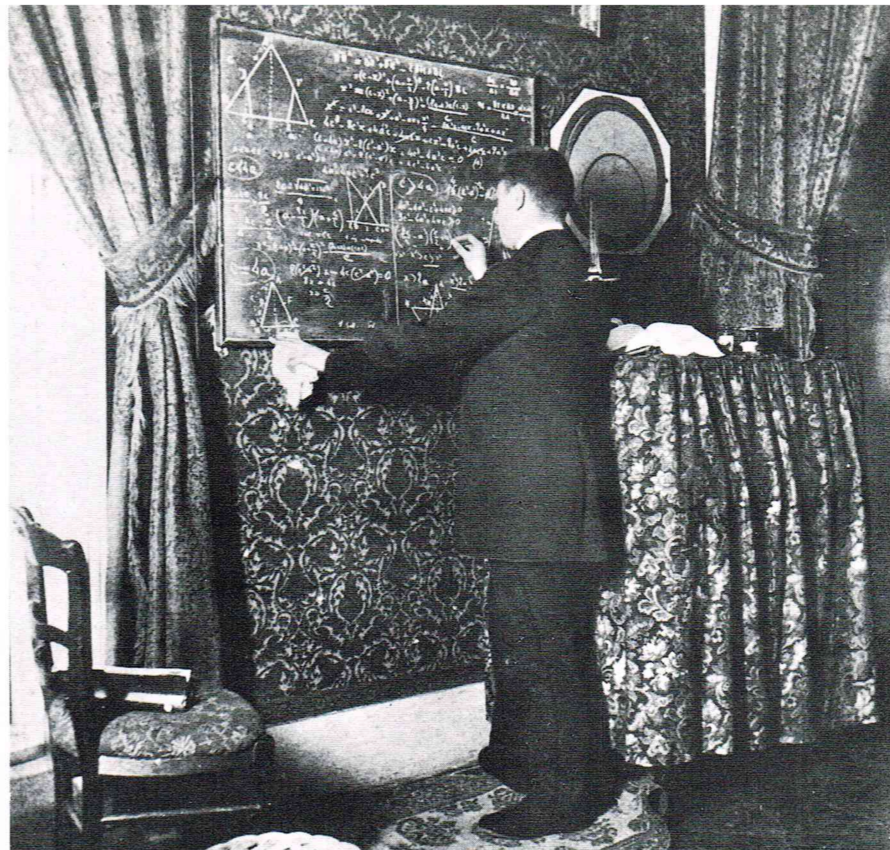
On a enseigné les mathématiques bien avant l'institution de l'école (au sens moderne), et réfléchi sur les difficultés de cet enseignement bien avant que n'apparaisse le souci de savoir s'il était ou non structurellement conforme à la mentalité de « nos enfants ». Aussi n'est-il peut-être pas étonnant qu'une approche pédagogique et didactique retrouve un problème qui, chez Platon, est rapporté aux fondements des mathématiques : au lieu d'escamoter le rapport entre un fait de langage tel que « 2 et 2 font 4 » et un fait mathématique tel que « $2 + 2 = 4$ », Platon s'acharne, pour ainsi dire, à élaborer cette difficulté. Socrate dit dans le *Phédon* être fort loin d'avoir compris l'addition et de pouvoir dire, lorsqu'à une unité on ajoute une unité, « si c'est l'unité à laquelle cette dernière a été ajoutée qui est devenue 2, ou si c'est l'unité ajoutée et celle qui a reçu cette adjonction qui, du fait même de cette adjonction de l'une à l'autre, sont devenues 2 ! » C'est, en effet, pour Socrate, un sujet d'étonnement « que, lorsqu'elles étaient, chacune, à part l'une de l'autre, chacune des deux visiblement unité et qu'alors il n'y eût pas de 2 ; et que, une fois qu'elles se sont rapprochées, il n'a fallu, paraît-il, pour faire qu'elles devinssent deux, d'autre cause que leur réunion par voie de mutuelle juxtaposition ». Voilà un étonnement qui ne serait plus de mise aujourd'hui dans un enseignement élaboré, mais tout à fait à sa place dans une véritable initiation, qui néglige forcément le détour de la méthode axiomatique abstraite.

C'est pourquoi la didactique des mathématiques est également, de manière à la fois spontanée et nécessaire, une réflexion sur la rigueur indispensable dans cet enseignement. — Du reste, répondant aux besoins de la société, l'enseignement des mathématiques s'est, à partir de la Révolution (et avec la création des « Écoles centrales », notamment), généralisé jusqu'à constituer aujourd'hui le centre de gravité de tout enseignement. Si, depuis le Moyen Age, on a successivement accordé de l'importance à la grammaire dans la mesure où elle semblait constituer le vestibule de la logique, puis à la langue en général et aux humanités parce qu'elles formaient le bon sens, il devient progressivement de plus en plus évident que la prépondérance doit être accordée aux mathématiques pour les mêmes raisons, les mathématiques n'étant dans leur partie abstraite, comme l'avait très bien exprimé A. Comte (*Cours de philosophie positive,* Hermann, 1975, p. 64), qu'une « immense extension admirable de la Logique naturelle à un certain ordre de déductions ».

L'importance du manuel au siècle des lumières

A vrai dire, le désir d'exposer rigoureusement les mathématiques fut d'abord, partiellement, un besoin didactique. C'est l'enthousiasme pédagogique du siècle des lumières qui engendre ces traités de mathématiques élémentaires, suscités par la création de nouvelles universités et de nouvelles chaires de mathématiques. Ce sont les auteurs de ces manuels, tant ridiculisés par un grand génie comme Gauss sous le nom de « Compendien-Schreiber », qui expriment d'abord cette volonté de rigueur. Parmi ces faiseurs de manuels, le plus fameux, A. G. Kästner, illustre admirablement la fonction de l'enseignement et de son organe essentiel, le *manuel,* très commode pour propager un savoir qui serait autrement resté l'exclusivité des spécialistes ! L'exemple de Kästner nous explique le mystère de la longue alliance de la littérature et des mathématiques, de la géométrie et du talent d'écrivain ; ainsi qu'ironisait Gauss, Kästner était « mathématicien parmi les poètes et poète parmi les mathématiciens ». Et c'est sans doute là l'origine de ces « exposés décoratifs » (Godement) qui permettaient, il n'y a pas si longtemps, à certains d'être candidats à l'Académie des sciences et à l'Académie française ! Mais c'est aussi et surtout l'origine de la nécessité, alors très relative, d'écrire pour les débutants des exposés aussi explicites que possible, où l'on essaie de tout démontrer et de tout définir. Par ses manuels, Kästner a permis à tous ceux qui ne pouvaient écouter un maître d'apprendre cependant les mathématiques. Le savoir mathématique s'est mis à circuler à une échelle





▲ *A gauche,*
A. de Villeneuve
enseignant la géométrie à ses élèves arpenteurs.
A droite, un étudiant travaillant à son tableau dans sa chambre en 1891.

plus populaire, conformément à la demande du rationalisme des lumières, expansif, prosélyte et intellectuellement réformateur. La création de nouvelles universités aux XVII[e] et XVIII[e] siècles s'accompagne ainsi d'une diffusion nouvelle des instruments du savoir. L'esprit géométrique alors à la mode descend des cimes vers le public cultivé, grâce au cours de Joseph Sauveur (1653-1716) où, à Paris, les gens se pressent, et grâce aux universitaires, tel Kästner, à Göttingen, qu'écoutent aussi ceux qui voient dans la science le seul moyen d'ascension sociale à leur portée.

Dans une lettre à Maupertuis, Kästner reconnaît lucidement que, s'il est incapable « d'éclairer le monde comme les grands génies... il sait au moins profiter de leurs lumières et les communiquer à d'autres qui, sans cela, n'en auraient peut-être tiré aucun autre avantage ». Bref, on n'en est plus à commenter Euclide ; on réfléchit sur la façon d'exposer avec rigueur les mathématiques.

Bernard Bolzano (1781-1848) développe une *théorie de l'exposition,* réflexion critique et méthodique inspirée par l'apparition et la popularisation des manuels depuis Kästner. B. Bolzano a justement appris les mathématiques dans les livres de Kästner, qui n'a pas inventé grand-chose, mais s'est évertué à prouver des évidences et à justifier toutes ses démarches, « à démontrer ce que d'autres passent allègrement sous silence, c'est-à-dire... à éclairer le lecteur sur la raison qui sous-tend un de ses jugements », note Bolzano, qui avoue par ailleurs en avoir été assez impressionné pour se mettre aux mathématiques (B. Bolzano, *Lebensbeschreibung,* Sulzbach, 1836, p. 19). Mais dans son propre effort de rendre les mathématiques accessibles et absolument claires à tout débutant, Bolzano découvre qu'il faut faire la *théorie de leur langage* et *expliciter la logique* qu'elles véhiculent dans leurs raisonnements. C'est ainsi qu'il est amené à réfléchir sur la notion de « proposition », de « démonstration », sur la nécessité de soumettre l'usage des symboles à des règles à la fois précises et explicites, à vouloir que l'effort d'explicitation aille jusqu'à réaliser la maxime leibnizienne « qu'il est bon de chercher les démonstrations des axiomes eux-mêmes », contrairement à l' « intuitionnisme » naïf ambiant dont un Ladomus s'était fait le défenseur dans ses *Contributions à la méthodique de la mathématique pure,* reléguant dans une discipline *extérieure* aux mathématiques l'exigence de *fonder clairement* ce qu'on avance. En considérant, au contraire, que cette exigence doit être *interne* aux mathématiques, Bolzano est conduit à refuser la philosophie mathématique de son époque, celle de Kant, et à critiquer la notion de jugement mathématique, dont il souligne le caractère analytique, c'est-à-dire le fait qu'il n'est subordonné à aucune autre discipline sinon justement la logique mathématique.

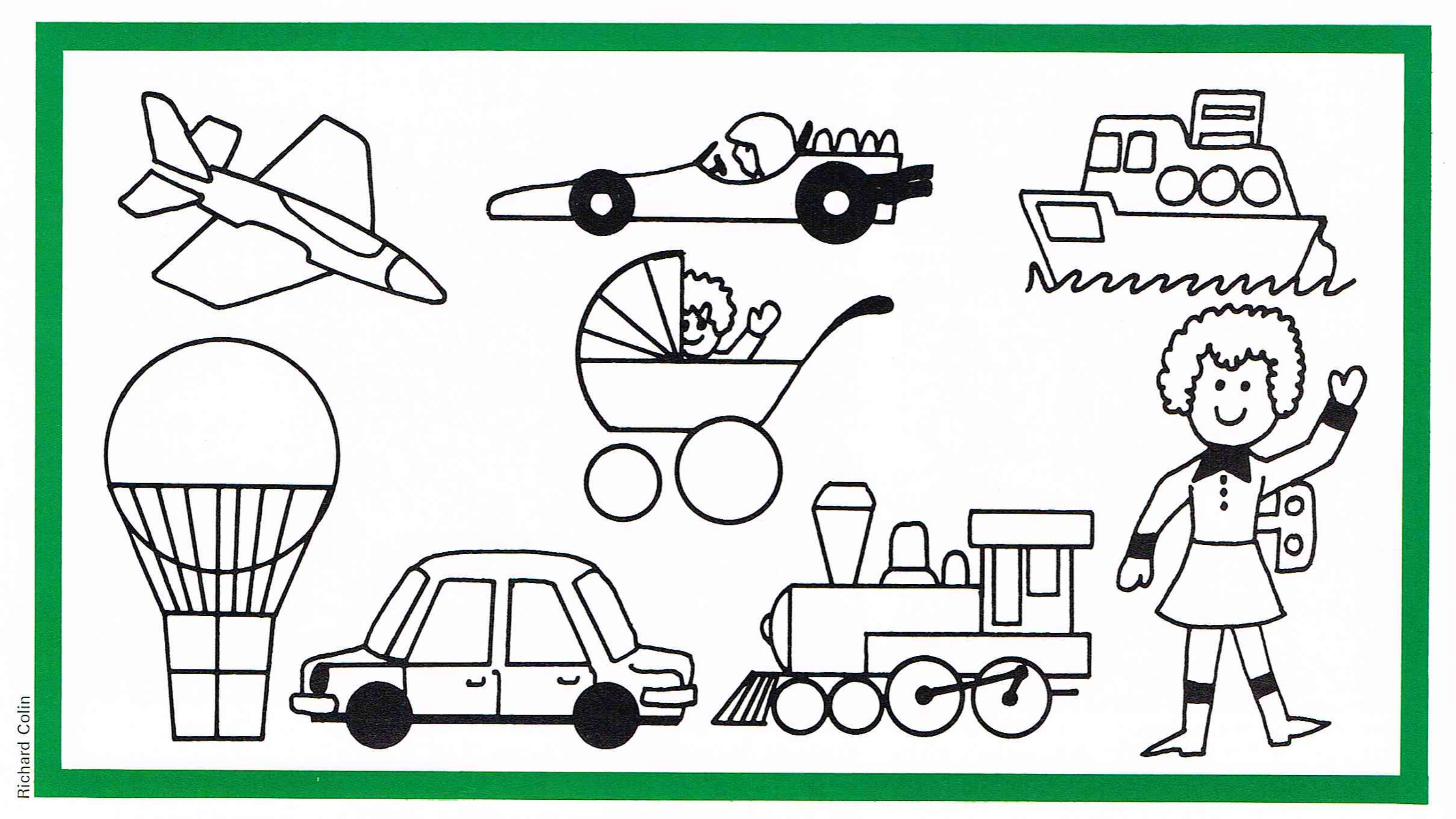
Le lien au XIX[e] siècle entre le souci de l'exposition et les exigences de rigueur logique

La réflexion sur la méthodique mathématique est donc liée à l'exigence de didactique persuasive et de rigueur dans l'exposition (en France, c'est Develey qui s'était fait l'écho, dans ses *Essais de méthodologie* [Paris, 1831], de ces préoccupations) ; elle sera bientôt suivie d'une réflexion sur les fondements logiques des notions élémentaires des mathématiques, et le point remarquable pour notre propos, c'est que cette réflexion ne cesse d'être liée à un souci pédagogique.

En effet, lorsque Richard Dedekind réfléchit aux fondements de la théorie des nombres, ce n'est pas seulement parce que celle-ci s'était révélée comme la partie essentielle des mathématiques, mais aussi en vue de rendre possible un enseignement rigoureux. Cela impliquait de renverser la priorité cartésienne de l'évidence sur la certitude et de satisfaire la nécessité de « convaincre » avant celle d' « éclairer » : d'abolir, donc, le recours à l'intuition géométrique dans la démonstration de théorèmes d'analyse. Tout cela, Dedekind en a fait l'expérience dans son enseignement, quand il était professeur à Zürich : il lui fallait, alors, expliciter les bases d'un savoir qui, pour être compris de tous, ne devait appartenir à personne, devait donc être dépouillé de tout arbitraire ; en abordant sans préjugés les problèmes qui se posent à la base, Dedekind élargit l'élucidation acquise aux sommets.

Mais l'arithmétique, science élémentaire, était aussi du ressort de l'enseignement secondaire. Les manuels de ce niveau attirent l'attention de G. Frege, l'initiateur de la logique mathématique moderne, qui, avant sa fameuse *Idéographie* (1879), ne dédaigne pas la tâche de les examiner. Il critique sévèrement le manuel de H. Seeger, *Die Elemente der Arithmetik,* où l'écolier n'est supposé mettre en œuvre que sa mémoire pour retenir les lois de l'arithmétique. Frege avait, du reste, été appelé à remplacer Karl Snell, dont il a retenu, comme il l'écrira plus tard, « le principe... qu'en mathématiques tout devrait être aussi clair que $2 + 2 = 4$ », *à condition que* « $2 + 2 = 4$ » *soit clair !* Méditant cet avertissement, Frege s'aperçoit que la « rigueur logique » n'est pas naturelle, qu'elle est le couronnement d'un travail mathématique qui se révèle peu susceptible, en tant que tel, d'être assimilé par des têtes mathématiquement vierges.

Bien sûr, les problèmes posés par l'enseignement des mathématiques n'ont pas, seuls, déterminé l'évolution de la réflexion sur les fondements des mathématiques. Mais ils sont ancrés dans le devenir des mathématiques du XIX[e] siècle, souvent de manière indirecte, dans les réformes consécutives à de nouvelles perspectives socio-économiques, et parfois de manière directe, puisque l'exposition doit être de nature à éviter tout obstacle à la


Richard Colin

▲ ***Colorier parmi ces jouets ceux qui roulent : exemple simple d'un exercice mal posé (repris du livre de Stella Baruk,*** **Échec et Maths*).***

compréhension d'une discipline en principe intelligible sans recours extérieur ni antérieur. Ce n'est donc pas un hasard si l'on distingue aujourd'hui entre l'analyste ou l'algébriste, qui peut tout ignorer de la logique mathématique, et l'enseignant qui a besoin au contraire d'une formation logique assez poussée (voir Daniel Lacombe, *Didactique des disciplines,* bulletin n° 2, janvier 1974, p. 4).

Pour l'enseignant soucieux de logique, la première nécessité est de prendre au sérieux le caractère logique des mathématiques; on ne saurait trop recommander aujourd'hui, où on initie de petits enfants à la théorie des ensembles, de se méfier des intuitions, des illustrations hasardeuses, des idées courantes sur de prétendus rapports entre les opérations concrètes effectuées par les petits de l'homme et les raisonnements mathématiques, qui en exprimeraient d'une certaine façon l'intériorisation. Outre les difficultés propres à cette métaphore de l'assimilation, de la digestion et de l'ingestion, on ne voit pas comment interpréter en termes de *manipulation* des propositions de logique, ni comment un travail, souvent considérable, d'élaboration logique peut se réduire à l'explicitation d'actes intériorisés ! On ne saurait trop recommander aussi d'éviter le recours aux « devinettes ». Bolzano déjà faisait une remarque qu'on aurait mille occasions de répéter aujourd'hui ; il observait que si on ne lui indique pas d'avance l'énoncé qu'on veut lui faire démontrer, l'apprenti mathématicien ne voit pas le but pour lequel telle construction ou tel chemin démonstratif s'impose. Par exemple, dans la démonstration traditionnelle du théorème de Pythagore, qui peut donner au débutant des raisons acceptables pour justifier le tracé de telles lignes, celles-ci plutôt que d'autres, si on ne lui a pas préalablement indiqué qu'il est utile de considérer le rapport entre les aires des carrés formés sur les côtés du triangle ? Sans doute, au lieu de désigner brutalement la voie à suivre, on pourra traiter des problèmes analogues et plus simples qui serviront de repères à l'élève.

Voici l'exemple simple d'un exercice mal posé, où l'on demandait aux enfants de colorier, parmi les jouets, *ceux qui roulent,* et qui déchaîna un drame à cause de l'avion dont on ne voit, sur le dessin, que les ailes et le fuselage, ce qui n'empêcha pas certains enfants de penser qu'il avait besoin de roues pour décoller !

La question est, en fait, posée de manière à renforcer les interférences, que les enfants n'ont de toute façon que trop tendance à laisser s'installer, entre les objets dessinés et ceux qu'ils utilisent réellement dans leurs jeux, ce qui ne leur facilite pas la tâche de saisir le caractère abstrait du *schéma* et la nécessité de l'observer pour lui-même. La représentation des jouets n'a de sens que par rapport à l'intention dans laquelle elle a été faite : qu'on colorie seulement les dessins d'objets où figurent des roues; « si c'est ça qu'il [le maître] veut, il n'a qu'à le dire », remarquent les enfants auxquels on explique après coup cette intention !

Le dire ? Ce n'est pas toujours facile, comme le savait bien Socrate, qui, après avoir vainement tenté de mettre l'esclave de Ménon sur la voie de la solution de la duplication du carré, lui dit enfin : « Essaie de nous répondre avec exactitude. Et si tu ne peux pas nous dire le nombre, *montre-le nous* » ; où un lecteur averti comprend que le côté du carré cherché ne peut être exprimé par un nombre ; il est « inexprimable », irrationnel, indication qui devrait suffire à faire penser à la diagonale du carré donné au départ. Il faut pour ainsi dire s'être déjà heurté à l'incompréhension des élèves pour chercher une formulation plus précise pour l'énoncé du problème posé – voir les remarques de Pierre Samuel sur « l'apparente objectivité de la sélection » par les examens dans *Pourquoi la mathématique ?* 10/18, pp. 158-170.

Mais si certains éléments implicites ne demandent qu'à être dits pour que tout s'éclaire, il en est que seule une connaissance de la *logique mathématique* permet d'éviter. Celle-ci est née justement, *avant* la découverte des fameux paradoxes de la théorie des ensembles, de la nécessité d'expliciter les principes du raisonnement mathématique et de formuler précisément certains faits d'analyse que l'ignorance de l'usage des quantificateurs laissait dans une grande confusion. On sait, par exemple, l'importance de la place des quantificateurs successifs qui interviennent dans les définitions de la continuité et de la convergence, et qui offrent un moyen à la fois simple et sûr de distinguer la continuité ou la convergence uniforme de la continuité ou de la convergence simple. D'où la nécessité d'introduire aussi tôt qu'on le peut des notions de logique et d'éviter de sous-entendre les quantificateurs à une époque où nul n'enseigne plus le « carré logique » aristotélicien, grâce auquel on pouvait éviter les confusions entre le contraire d'un énoncé et sa négation.

Cependant, la mise au point d'une didactique « honnête » maintient toujours l'enseignement mathématique dans l'ornière de la méthode, de l'intelligence de l'acquis, donc de l'apprentissage et du « dressage ». Il n'échappe pas au risque de réduire l'enseignement au montage correct d'automatismes efficaces. Cela pose deux problèmes :

d'une part, que convient-il d'exiger de l'apprenti mathématicien? Un apprentissage effectif, qui fait donc plus ou moins appel à sa mémoire, ou une capacité d'imaginer des médiations, c'est-à-dire une intelligence générale indépendante de l'apprentissage? D'autre part, on peut se demander si un véritable enseignement des mathématiques ne serait pas plutôt une initiation à la découverte, découverte mathématique et non découverte des mathématiques. Or la solution de ce dernier problème dépend de la conception qu'on a des mathématiques; discipline spéculative, où il vaudrait donc mieux apprendre à créer, un peu comme en philosophie on apprend — en principe — à philosopher plutôt que le contenu des doctrines philosophiques; ou discipline utile, soit que l'on croie y résoudre les questions significatives par leurs applications, soit qu'on prétende s'inscrire dans une tradition de problèmes prestigieux. Quelles que soient les réponses à ces questions, on voit ce qu'elles mettent en jeu : comment, c'est-à-dire quelles mathématiques enseigner?

Le contenu de l'enseignement - Des programmes scolaires

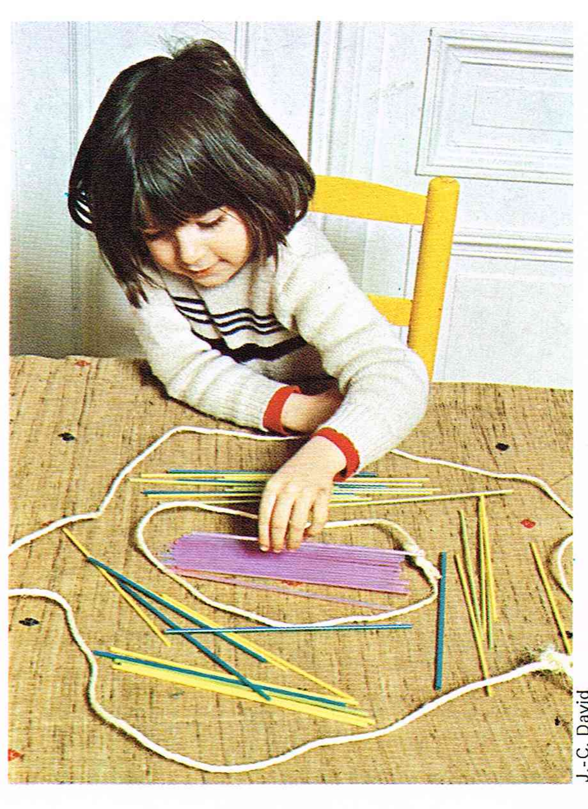
La question du contenu de l'enseignement des mathématiques se pose dans la mesure où les nouveaux développements des mathématiques ont exigé un changement de structure de l'enseignement primaire et secondaire, qui en harmonise les programmes avec ceux de l'enseignement supérieur.

Certains auront entendu la fameuse déclaration de ce professeur d'université qui commençait l'année par ces mots : « Il y a une seule façon de comprendre le cours que je vais vous faire, c'est d'essayer d'oublier tout ce que vous avez appris au lycée »! Derrière l'incohérence ou la discontinuité que dénonce cette boutade, se cache un problème philosophique : celui que Frege désignait comme celui des « définitions par morceaux », dont il dénonçait l'usage courant, surtout depuis les extensions successives des nombres. Concepts et opérations sont ainsi définis pour un domaine restreint, celui des nombres entiers positifs, par exemple; puis, avec l'introduction d'éléments nouveaux, les nombres négatifs, fractionnaires, irrationnels, etc., on redéfinit exprès ce qu'on avait déjà défini, mais de façon à englober les opérations et les concepts propres à régir les éléments nouveaux. Ainsi les symboles n'ont pas de signification « définitive », et les concepts paraissent incohérents, puisqu'on demande à l'élève d' « oublier » dans le secondaire qu'il était interdit d'écrire « $3 - 5$ », et à l'Université qu'il le fut de poser $\sqrt{-5}$. Logiquement, cela signifie que la valeur de vérité d'un énoncé d'un langage dépend des structures qui constituent les réalisations de ce langage (voir la définition des concepts de « réalisation » et de « modèle » d'un langage formel dans le chapitre *Logique*).

Rigoureusement, il faudrait donc disposer du concept d'un anneau quelconque K, et expliciter la quantification sous-jacente, afin qu'on voie que « y est un carré dans K » signifie « il existe x dans K et le carré de x est égal à y ». Autrement dit, un enseignement homogène d'une mathématique sans ruptures se fonde sur la méthode axiomatique, qu'on suppose aussi familière que les principes élémentaires de la déduction aux étudiants entrant à l'Université. C'est pour que cette supposition corresponde à une réalité qu'ont été introduites les mathématiques « modernes » dans l'enseignement primaire et secondaire. Cette innovation correspond à l'exigence de limiter la prédominance de la géométrie et, par là, du recours facile à l' « intuition » ou visualisation concrète et singulière d'objets susceptibles d'une définition plus abstraite et plus générale. La pensée abstraite doit, certes, au niveau de l'enseignement secondaire, laisser sa place à une certaine visualisation, par exemple dans l'étude et la représentation graphique des fonctions par leurs courbes; mais il importe que celle-ci se déploie dans les limites clairement déterminées par un cadre logique et axiomatique.

C'est ainsi que M. Dieudonné avait, il y a plus de quinze ans (en 1959, à l'occasion d'une session d'études consacrée à la réforme de l'enseignement des mathématiques par la Direction des affaires scientifiques), conçu un programme *moderne* d'enseignement des mathématiques, dont voici les grandes lignes :

— Avant 14 ans, on se bornerait à un travail pour ainsi dire expérimental sur l'algèbre et la géométrie plane, en ne soulignant les déductions logiques qu'aux endroits où c'est possible. La part du lion est faite à l'algèbre, dans la mesure où la géométrie elle-même s'organise autour des notions fondamentales de symétrie, translations, produit de transformations, etc. On n'omettrait pas d'introduire le langage adéquat, c'est-à-dire le langage commun à toutes les branches des mathématiques actuelles, et si possible, de développer les règles ordinaires de l'arithmétique à partir des axiomes de Peano, dont l'enseignant aura montré la nécessité aux élèves en les faisant réfléchir sur notre acceptation de la validité des lois de l'arithmétique pour des nombres inaccessibles à notre intuition.


J.-C. David

◀ *Faire comprendre les concepts d'ensemble et de nombre en s'appuyant sur des schémas, mais sans mutiler totalement ces notions de leur caractère d'entités abstraites.*

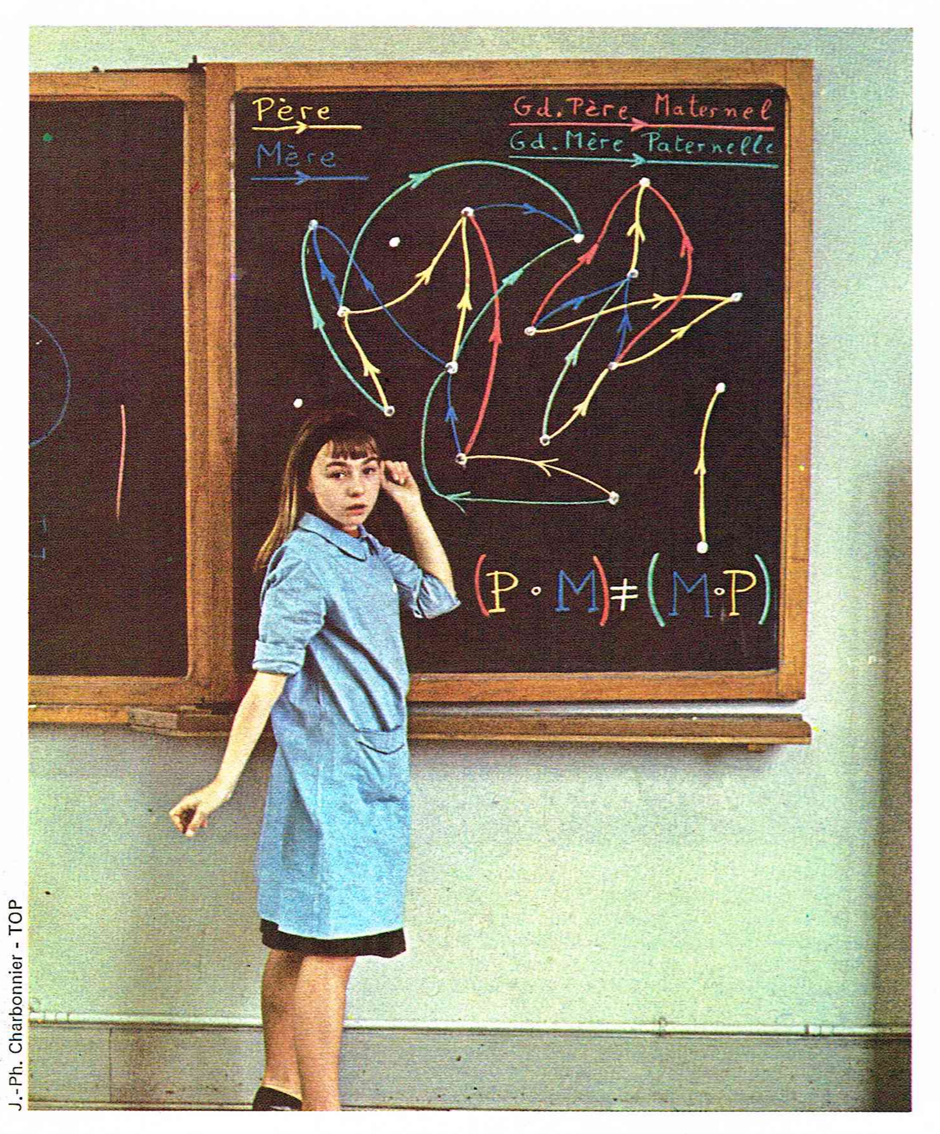
— A 14 ans, on introduirait la notion de courbe représentative d'une fonction et différentes méthodes d'approximation. On peut aussi présenter, non pas une *définition,* mais une *étude axiomatique* des nombres réels, et donner, comme résultats, ultérieurement démontrables, l'existence du maximum, et le théorème de Bolzano sur l'existence des racines.

— A 15 ans, il est possible de formuler les axiomes exprimant que, pour l'addition, un espace vectoriel est un groupe commutatif, en prenant soin de développer les conséquences logiques tant du point de vue algébrique que du point de vue géométrique, c'est-à-dire en donnant les conséquences avec les deux interprétations possibles, ce qui fournit l'occasion d'entrevoir le concept de *modèle*.

— A 16 ans, le décollage axiomatique ayant déjà eu lieu, on approfondirait l'étude des groupes de la géométrie plane, en définissant la notion de mesure des angles comme homomorphisme du groupe des nombres réels sur le groupe des rotations, et en introduisant les nombres complexes avec leur interprétation géométrique. L'aspect technique consisterait dans l'étude de la notion de fonction primitive et de la notion d'aire pour des domaines simples, avec des exemples élémentaires.

— A 17 ans, la géométrie à trois dimensions doit être présentée axiomatiquement, avec emploi des matrices et des déterminants d'ordre 3. Du point de vue technique, on montrerait l'utilisation des fonctions primitives pour calculer les volumes simples, et l'on introduirait les coordonnées polaires et la méthode de construction d'une courbe donnée par une équation en coordonnées polaires. On peut, encore, étudier les fonctions logarithme et exponentielle, sans démonstration d'existence mais en insistant sur le fait qu'il s'agit d'homomorphisme de groupes.

Voilà pour le secondaire. Mais comment y préparer les enfants plus jeunes? Les problèmes les plus délicats



J.-Ph. Charbonnier - TOP

▲ ***Selon la méthode préconisée par Georges Papy, une application des graphes au jeu des relations familiales. Si d'un point part une flèche bleue vers un second point et que de celui-ci soit issue une flèche jaune vers un troisième point, ce dernier est le grand-père maternel du premier; de l'un à l'autre, traçons directement une flèche rouge qui représente le produit de la relation*** M ***(comme mère) par la relation*** P ***(comme père). De même, une flèche jaune suivie d'une flèche bleue désigne la grand-mère paternelle, c'est-à-dire le produit*** P ***par*** M ***que nous désignons directement par une flèche verte. Comme le grand-père maternel est différent de la grand-mère paternelle, les enfants écrivent au tableau l'équation*** **(P · M) ≠ (M · P)** **[*noter la couleur différente des parenthèses*].**

se posent là, car on ne peut être rigoureux qu'au risque d'être inaccessible, quand on s'adresse à de petits enfants. Il s'agit alors moins de respecter les réquisitoires de la rigueur déductive que d'expliciter le mieux qu'on peut, sans fermer les portes à l'imagination et à l'initiative, et sans perdre de vue le but qui est d'arriver à ce que les mathématiques soient saisies comme une activité créatrice. Des essais intéressants ont été tentés dans cette perspective, parmi lesquels il faut citer d'abord la *Mathématique moderne* de Papy, dont l'ambition est de faire participer le débutant à la construction active de l'édifice mathématique à partir de situations simples, et de l'initier — dès l'âge de douze ans — aux éléments de la théorie des ensembles.

Moins connu en France, sans doute, est le bel effort de Patrick Suppes, grand mathématicien américain qui a consacré à la théorie abstraite des ensembles un traité fameux, mais qui a aussi conçu un programme destiné aux enfants des maternelles auxquels il s'agit essentiellement de faire comprendre les *concepts d'ensemble* et de *nombre*, en s'appuyant, certes, sur des schémas, mais sans mutiler totalement ces notions de leur caractère d'entités abstraites. Partant du concept d'un ensemble de choses comme *famille de choses* semblables entre elles (les oiseaux d'un même nid), ou différentes (une banane et un livre), nombreuses (les feuilles d'un arbre), moins nombreuses, réduites à une seule chose ou même à *rien*, il introduit le concept de nombre comme propriété d'un ensemble, en distinguant soigneusement dans la notation un ensemble, par exemple {crayon, balle} de son nombre N {crayon, balle}. Cette première initiation comporte les notions de plus grand, plus petit, plus long, plus court, plus haut, la notion d'égalité et la reconnaissance des formes géométriques semblables quoique de dimensions différentes, enfin l'apprentissage des nombres, de 0 à 3 dans un premier temps, puis jusqu'à 5. Le tout présenté sous forme de jeux dirigés pour susciter l'initiative dynamique et faciliter l'assimilation ultérieure des notions et des méthodes.

On retiendra de ces exemples qu'un bon enseignement des notions élémentaires repose sur une *analyse logique* de ces notions et du contenu mathématique (compter, ajouter, retrancher, etc.) qui s'ensuit, d'où l'exigence pour les futurs enseignants du primaire et du secondaire de se former aux principaux concepts de la logique mathématique.

Les programmes se concrétisent naturellement dans des manuels, et les bons manuels sont rares, si bien qu'ils apparaissent souvent comme l'expression d'une mathématique figée ou d'un enseignement dissocié de la recherche qui devrait pourtant le nourrir. Les pédagogues les ont, depuis un moment, tant critiqués qu'ils ont pratiquement disparu de l'enseignement au profit des *fiches*, fiches de notions et fiches d'exercices pour appliquer ces notions.

Ainsi, dès 1929, Freinet condamnait le manuel comme trop étriqué et trop scolastique et lui préférait le fichier, permettant un enseignement plus souple et faisant davantage appel à l'initiative tant du professeur que des élèves. Un ami de Freinet, R. Dutheil, a également attiré l'attention, dans un article de *l'École libératrice* (9 et 30 janvier 1932), sur l'arithmétique programmée de Winnetka, et les efforts suscités par ce genre de préoccupations aboutissent aux *bandes enseignantes*, autocorrectrices et complétées généralement par un test.

L'avantage des fiches est qu'elles permettent de réindividualiser la vie pédagogique, ce qui constitue le but de l'*enseignement programmé*. On a ainsi tenté une programmation de l'algèbre élémentaire, et il existe également une arithmétique programmée présentée sous forme de fiches d'information et d'exercices de contrôle. Bien qu'il ait favorisé l'étude de tous les chaînons utiles dans un apprentissage, l'enseignement programmé ne semble pas avoir tenu toutes ses promesses; en particulier, il n'évite pas l'écueil de la présentation collective, par exemple, devant une classe d'enfants dont la maturité ou la rapidité d'assimilation n'est pas nécessairement uniforme, ni n'élimine vraiment l'idée de « l'écolier moyen » dont le profil détermine la constitution des fiches.

Diversité des idées pédagogiques, multiplicité des méthodes, il y a de quoi rendre perplexes ceux dont la tâche est de former les futurs formateurs. « Comment concilier l'effort de l'éducateur sur lui-même pour s'ajuster à son rôle, et l'effort pour ajuster ses interventions à l'enfant et à son contexte social ? » se demandait J. Ferry. Plus on approfondit les problèmes pédagogiques, plus on s'aperçoit que, théoriquement, leur solution pose des exigences draconiennes dont les institutions actuelles ne permettent pas la réalisation harmonieuse. La situation est bien résumée dans cette remarque de R. Thom que nous citons : « Pour s'assurer que l'élève participe pleinement à la recherche en cours, il est nécessaire que le maître tienne compte à tout moment de ses réactions, afin de guider sa propre démarche et celle de son pupille. Ceci n'est guère possible, sous forme idéale, qu'en tête-à-tête... Dès qu'un maître a simultanément plusieurs élèves, il ne peut tenir compte des réactions souvent diverses de tous ses élèves, et il est amené à en négliger quelques-uns. Encore un pas de plus, et le souci d'efficacité le conduira à adopter une attitude de guide, et bientôt à revenir à l'enseignement *ex cathedra*. Aussi les efforts vers une pédagogie plus libre sont-ils nécessairement coûteux : ils demandent plus de maîtres et des maîtres mieux formés, à personnalité originale : la société ne pourra qu'imposer à ces efforts les bornes inéluctables du budget. »

Au total, la pédagogie des mathématiques est-elle pour le moment autre chose que le bilan des difficultés inhérentes à la notion même d'éducation ? autre chose que le constat du désarroi de l'éducateur qui, en principe, ne peut se contenter de tout organiser en fonction d'un système scolaire conventionnel, en négligeant les virtualités de ses élèves et les potentialités de changement de la société actuelle ? L'enseignant, pris en tenailles entre des exigences peu compatibles, est la victime d'une situation de crise, provoquée par l'effondrement des idéaux humanistes et religieux, et qu'on n'a pas encore surmontée.



▲ *Les plus anciens documents mathématiques de la civilisation égyptienne remontent à une époque très lointaine ; ces peintures murales ornant un tombeau royal, à Louxor, datant du XIIIe siècle avant J.-C., représentent des agriculteurs au travail dans un champ de blé et, en bas, des scribes effectuant des calculs.*

HISTOIRE DES MATHÉMATIQUES

L'historien des mathématiques se heurte aux difficultés de tout historien, du fait qu'il n'y a pas trace de tout ce qui s'est passé, que l'écrit lui-même n'est pas indélébile, que bien des éléments utiles à la connaissance historique ne furent consignés qu'après coup, après avoir traversé le flou du rapport oral, que les perspectives particulières, ou même les passions, ont souvent contribué à former une image plus conforme à la situation dans laquelle l'histoire était écrite qu'à la vérité, introuvable ou fragile. Tout se complique encore si l'on essaie de définir le sens de l'histoire des mathématiques du point de vue des mathématiques actuelles. Car, ou bien on discerne un lien entre les contenus actuels et passés, et on évite mal le risque d'appauvrir un contenu d'autant plus fuyant qu'il est dissocié de son contexte, ou bien on n'en discerne pas, et l'on se trouve face à des fragments auxquels il est malaisé de donner un sens en les rapportant à un contexte composé d'éléments hétérogènes. Un philosophe rigoureux pourrait trouver dans l'entreprise même de tenter une histoire des mathématiques un vice fondamental, celui d'être un intolérable mélange des genres. Aussi n'avons-nous pas la prétention d'écrire, dans l'espace qui nous est imparti, une histoire qui aura surmonté toutes ces difficultés ; mais ce sera assez de les circonscrire. Bien entendu, nous nous limiterons aux faits les plus prégnants, dans le cadre actuel de connaissances qui ne cessent de s'augmenter et de s'améliorer.

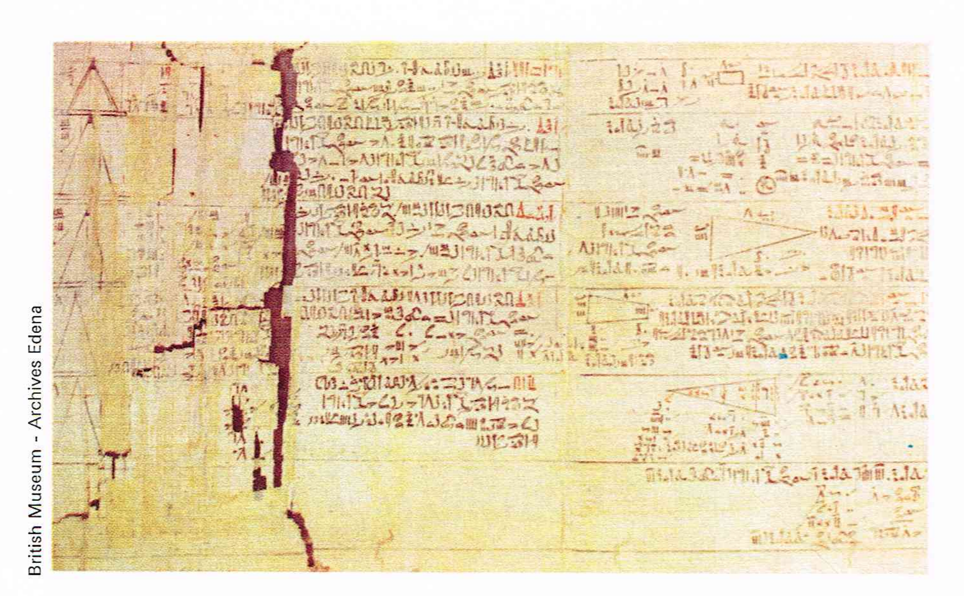
Les mathématiques orientales

Nous nous garderons de parler de « mathématiques préhelléniques » et laisserons entière la question de savoir dans quelle mesure la Grèce fut l'héritière de l'Égypte et de Babylone. Car rien ne démontre là cette continuité qu'on est tenté de supposer, sur le témoignage des Grecs eux-mêmes (Aristote ou Hérodote). On ne peut, en effet, étendre à toutes les périodes ce que Hilbert disait de la mathématique moderne, dont chaque moment pose des problèmes que l'époque suivante résout ou néglige comme stériles en leur en substituant d'autres. Mais, même si, avant elle, les méthodes primitives du compter et les nombreuses techniques ornementales qui utilisent des formes géométriques usuelles : points, lignes, carrés et losanges, cercles et spirales, portent la trace d'une pensée mathématique, c'est sans conteste aux antiques civilisations de l'Orient que revient le droit d'avoir inauguré cette science.

L'Égypte

Les sources dont nous disposons pour la connaissance des mathématiques égyptiennes ne sont pas très riches ; elles se limitent aux papyrus fragmentaires du Moyen Empire (1900-1800 avant J.-C.) : papyrus de Kahoum et de Berlin, aux textes des fameux papyrus Rhind et de Moscou, à un court manuscrit du British Museum, aux deux tablettes sur bois du musée du Caire, enfin, aux scènes de la vie égyptienne qui nous montrent l'importance des scribes, lorsqu'il s'agit par exemple de faire des mesures ou d'évaluer des récoltes.

Le plus important de ces documents est le papyrus Rhind (du nom du Britannique qui l'a acheté en 1858 à Louxor et légué au British Museum). Comme presque tous les autres documents mathématiques qui nous sont parvenus, il date de la période des Iksos, rois d'origine asiatique implantés dans le delta du Nil au XVIIIe siècle avant notre ère. Composé par le scribe Ahmès, ce papyrus



British Museum - Archives Edena

▲ *Le papyrus Rhind, d'après un traité égyptien de la XII[e] dynastie. On y reconnaît des problèmes de fractions et des calculs de volumes et de surfaces (British Museum).*

se présente comme une sorte de manuel. Le début du texte, prometteur, annonce qu'on va dévoiler l'essence de toutes choses. Mais, en fait, il s'agissait plutôt d'initier à un enseignement déjà codifié les scribes, devenus une sorte de caste indispensable à l'administration pharaonique ; or, prescriptions utilitaires et recettes techniques suffisent à ce but, et c'est précisément la forme qu'ont les problèmes (au nombre de 84, et de complexité croissante) du papyrus Rhind, ainsi que ceux (au nombre de 25) du papyrus de Moscou. Les énoncés arithmétiques ou géométriques qu'on y trouve appartiennent donc à des *disciplines auxiliaires* et *utiles*.

L'arithmétique égyptienne

Les nombres s'écrivaient (sous forme de hiéroglyphes) dans un système décimal, avec des signes spéciaux pour les unités, les dizaines, les centaines, les milliers, les dizaines de mille, les centaines de mille et les millions. En revanche, il n'y avait pas de signe pour zéro, bien qu'il soit pour ainsi dire indiqué par une place vide (*fig.* 1).

Chacun de ces signes est répété autant de fois qu'il est nécessaire pour exprimer le nombre désiré d'unités, de dizaines, de centaines, etc., les chiffres les plus élevés étant écrits avant les autres (*fig.* 2).

L'addition est une opération simple dans ce système ; mais il n'en va pas de même pour la multiplication et la division qui ne se font directement que par 2. Le problème n° 32 du papyrus Rhind conduit ainsi la multiplication de 12 par 12 (*fig.* 3).

Le résultat final est donc la somme de 4 × 12 et de 8 × 12 (= 144), les résultats à additionner ayant été cochés pour être distingués de ceux qui restent accessoires.

fig 1

1	|
10	∩
100	
1 000	
10 000	ou
100 000	
1 000 000	

Richard Colin

fig. 2

2 375 486 =

Richard Colin

I	∩II	1 (fois) : 12	
II	∩∩IIII	2 (fois) : 24	
IIII	∩∩∩∩IIII IIII	4 (fois) : 48	/
IIII IIII	∩∩∩∩∩∩∩∩∩ IIIIII	8 (fois) : 96	/

fig. 3

Richard Colin

▲ *Dans l'arithmétique égyptienne, les nombres s'écrivaient (sous forme de hiéroglyphes) avec des signes spéciaux pour les unités, les dizaines... (fig. 1). Chacun de ces signes est répété autant de fois qu'il est nécessaire, pour exprimer le nombre désiré d'unités (fig. 2).*

La division est une multiplication inversée. Le n° 69 du Rhind demande : « Additionne, en commençant par 80, jusqu'à ce que tu obtiennes 1 120 », ce qui veut dire qu'on cherche combien de fois 80 est contenu dans 1 120. L'opération 1 120 : 80 = 14 va être conduite comme une multiplication :

	1	80
/	10	800
	2	160
/	4	320
	14	1 120

On additionne les résultats cochés.

Quand le dividende n'était pas exactement divisible par le diviseur, on recourait aux fractions. Des fractions comme $\frac{1}{2}, \frac{1}{3}, \frac{1}{4}$ qui ont l'unité pour numérateur, sont usuelles et s'expriment par le signe qui signifie « part de », par exemple :

$$= \frac{1}{12} \qquad = \frac{1}{30};$$

mais $\frac{2}{3}$ a (seule) un signe spécial :

Les fractions n'ayant pas l'unité au numérateur sont décomposées ; par exemple :

$$\frac{2}{7} = \frac{1}{4} + \frac{1}{28} \quad \text{ou} \quad \frac{2}{5} = \frac{1}{3} + \frac{1}{15}$$

Sur les fractions, le calcul se fait comme sur les nombres entiers, par le *procédé de la duplication,* mais on suppose parfois connues certaines relations immédiates, telles que :

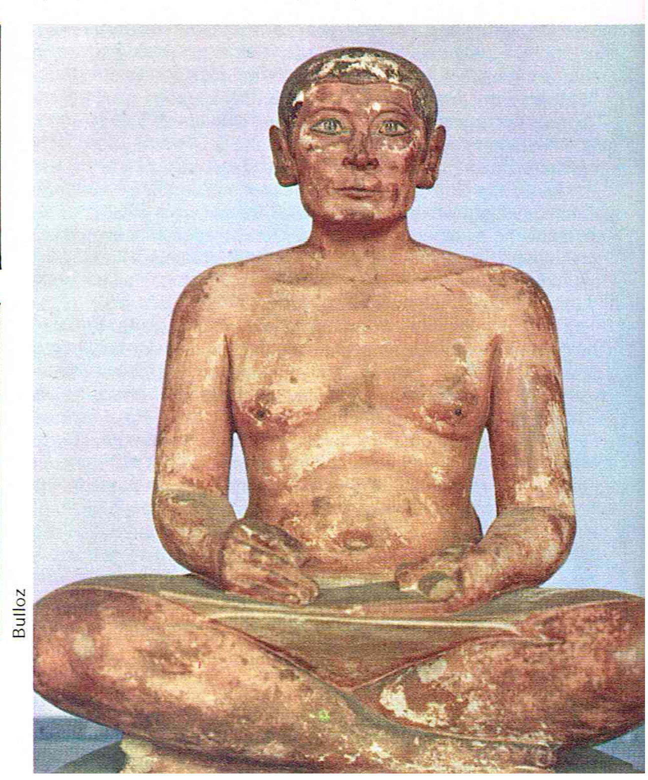
$$\frac{1}{6} + \frac{1}{6} = \frac{1}{3} \qquad \frac{2}{3} + \frac{1}{2} = 1 + \frac{1}{6} \qquad \frac{1}{3} = \frac{1}{4} + \frac{1}{12}$$

qui sont utilisées comme des règles dans le calcul des fractions les plus courantes (1/2, 1/3, 1/6).

Pour faciliter certains calculs, le papyrus Rhind donne une table de décomposition des fractions du type $2/n$, pour n allant de 3 à 101, mais les techniques employées ne semblent pas se ramener à une méthode unique ; cela constitue toutefois un des éléments les plus importants du contenu de ce papyrus.

En bref, ces calculs contiennent beaucoup de problèmes qui sont pour nous du ressort des équations linéaires.

Par exemple, l'expression $x + \frac{x}{7} = 19$ semble avoir un intérêt purement théorique et suggère la question de savoir si les Égyptiens ont, comme les Babyloniens à la même époque, utilisé le calcul algébrique. Mais on se trouve là



Bulloz

devant une interrogation à laquelle on ne peut répondre.

Le n° 40 du Rhind mérite aussi d'être cité, car il revient à partager 100 pains entre 5 hommes de façon que les parts forment une progression arithmétique ; l'idée de progression arithmétique reste, bien entendu, très implicite dans les essais successifs qui aboutissent à la solution. Celle de progression géométrique est également à l'œuvre dans le problème n° 79 qui demande la somme de ce que contient un domaine composé de 7 maisons, possédant chacune 7 chats, chacun des chats tuant 7 souris dont chacune mangeait 7 grains d'orge, alors que chaque grain aurait produit 7 boisseaux. Dans tous les cas, l'apparence pragmatique évidente fait ressortir ce qu'on pourrait appeler la profondeur mathématique implicitement enfouie dans les outils utilisés.

La géométrie égyptienne

Pour les Grecs, l'Égypte était le berceau de la géométrie, qui leur semblait une science assez ancienne pour que Platon fît dire à l'Égyptien du *Timée* (s'adressant à Solon) : « Vous autres Grecs, vous êtes toujours des enfants, vous n'avez aucune science blanchie par le temps. » Sans doute, la géométrie égyptienne n'est pas une science, au sens où les Grecs entendaient ce terme, mais plutôt un calcul appliqué, une pratique de l'évaluation des surfaces. C'est d'Hérodote que nous tenons l'idée que cette évaluation était indispensable à l'établissement de l'impôt. Mais nous pouvons constater directement l'origine empirique de cette géométrie dans certains problèmes du Rhind, par exemple dans celui (n° 51) où l'on demande de calculer la surface d'un triangle de 10 verges de hauteur et 4 verges de base en prenant la moitié de 4 pour en faire un rectangle et en multipliant 10 par 2. C'est donc en prenant la surface du rectangle construit sur la moitié de la base du triangle (et ayant comme longueur la hauteur du triangle) qu'on a la solution cherchée. Tout semble indiquer, justement, que le rectangle est la figure de base à partir de laquelle on calcule l'aire du triangle rectangle, puis du triangle isocèle, enfin celle du trapèze (*fig.* 4).

Le papyrus Rhind nous donne une approximation de l'aire du cercle, qui est considérée égale au carré des $\frac{8}{9}$ du diamètre, ce qui revient à l'évaluation suivante pour π :

$$\pi \cong 4\left(\frac{8}{9}\right)^2 = 3{,}160\ 49...$$

Ce calcul étonnant de π est à l'origine de nombreuses spéculations. Gaston Milhaud l'attribuait ainsi à l'efficacité d'un procédé qui ne nous serait pas parvenu, ce qui implique l'idée, parfois défendue, d'une science égyptienne vivante, par rapport à laquelle les recueils de problèmes que nous connaissons ne représenteraient que des formes cristallisées, figées et comme pétrifiées. Cette impression se renforce quand on remarque que le papyrus de Moscou livre la formule exacte pour calculer le volume du tronc de pyramide : $V = \frac{h}{3}(a^2 + ab + b^2)$, laquelle, comme le note Wan der Waerden, ne peut être calculée de manière purement empirique et résulte probablement de quelque considération théorique (*fig.* 5).

Malheureusement, on en est réduit à des conjectures, bien que l'on puisse imaginer des *constructions* rendant compte des formules sans induire des hypothèses dont le niveau dépasse celui qui nous apparaît aujourd'hui : rien ne permet de définir la réelle profondeur des formes condensées dans lesquelles nous sont parvenues les solutions des problèmes de la mathématique égyptienne. On retiendra donc que cette mathématique nous est parvenue à travers une collection d'*exemples* de problèmes, sans qu'il y soit jamais question de *démonstration*. En outre, il ne semble pas qu'elle ait connu une *évolution* ou un *progrès* quelconque durant tout ce temps où elle fut principalement l'outil des architectes, des arpenteurs et des maîtres calculateurs. Mais ni le calcul, ni l'arpentage ne suffisent, dit-on, à créer une véritable « science ».

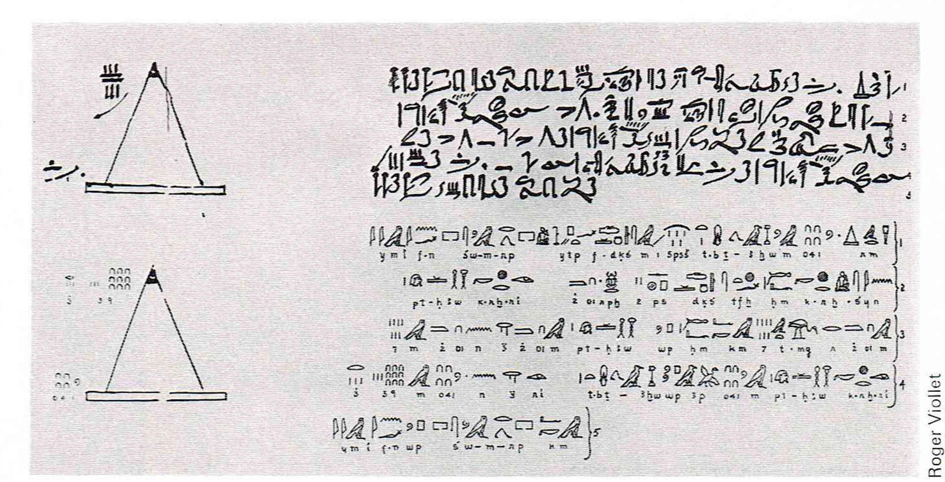
La Mésopotamie

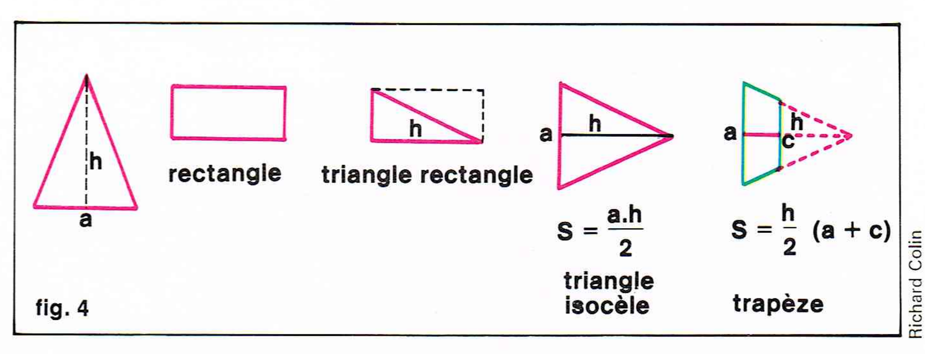
Le pays qui s'étend entre le Tigre et l'Euphrate développait, depuis des millénaires également, une civilisation agricole et urbaine. Mais, avec l'arrivée des Sumériens, peuple aux origines inconnues, venu du golfe Persique, s'épanouit une culture originale qui va durer jusqu'à l'aube de l'ère chrétienne. Vers 2200 avant J.-C., les Babyloniens, venus du nord du pays, soumettent les Sumériens et étendent leur empire jusqu'à la Méditerranée. Ils fondent la dynastie akkadienne, avec Babylone pour capitale, mais les « vaincus ayant conquis les vainqueurs », la culture sumérienne continue de s'épanouir. Elle atteint son acmé durant la dynastie de l'illustre législateur : Hammourabi (vers 1780 avant J.-C.).

Un peu plus tard, se développe en Assyrie une autre civilisation remarquable, tributaire des Sumériens et des Akkadiens, mais douée également d'une puissante originalité. C'est elle qui dominera, en fin de compte, la région, car elle saura mieux se défendre contre les envahisseurs, contre les rivaux politiques de Babylone, et atteindra son apogée au VII[e] siècle. Aussi avons-nous de cette culture des traces relativement tardives, proches du début de l'ère chrétienne, bien qu'elle ne survive plus à ce moment que dans les temples où les textes cunéiformes résistent à l'invasion de la langue du Perse vainqueur.

La notation numérique

Ce degré d'abstraction atteint par la science suméro-akkadienne est attesté avant tout par l'apparition de l'*écriture cunéiforme,* qui implique les notions d'idéo-


Roger Viollet



Richard Colin

Tronc de pyramide de hauteur h et dont la base est un carré de côté a ; b est le côté du carré suivant lequel est faite la section de la pyramide. Empiriquement, on peut subdiviser le tronc de pyramide en 1 parallélépipède, 2 prismes et 1 pyramide plus petite. **fig. 5**

b

h

a

$V = \frac{h}{3}(a^2 + ab + b^2)$

Richard Colin

▲ En haut, problème de géométrie égyptien, antique, et sa traduction en caractères dénotiques. Ci-dessus, figure 4 : calcul des aires du triangle rectangle, du triangle isocèle et du trapèze à partir du rectangle, expliquant un problème du Rhind.

◄ Figure 5 : schéma dû à O. Neugebauer pour expliquer le calcul du volume du tronc de pyramide du papyrus de Moscou.

◄ Page ci-contre, à gauche, figure 3 : la multiplication de 12 par 12 telle qu'elle est conduite dans le papyrus Rhind. A droite, Le scribe accroupi (musée du Louvre, Paris).

Tableau I - Système de numération babylonienne

(0 :)	6 :	12 :	60 :	120 :
1 :	7 :	20 :	70 :	180 :
2 :	8 :	21 :	80 :	200 :
3 :	9 :	30 :	90 :	etc.
4 :	10 :	40 :	100 :	
5 :	11 :	50 :	101 :	

Dans la pratique : 60 : (1 ŠU = une soixantaine) – 100 : (1 ME = un cent) – 1.000 : (1 LIM = un millier).

Fractions : Système savant : 1/2 : (30/60) ; 1/3 : (20/60) ; 1/4 : (15/60) ; etc.
Dans la pratique : 1/2 : ; 1/3 : ; 2/3 : ; 5/6 : .

Richard Colin

▲ *Tableau I : la numération babylonienne.*

▼ *Figures 6, 7, 8 : exemples d'opérations arithmétiques babyloniennes.*

gramme et de valeur syllabique, et, en ce qui concerne les mathématiques, par une connaissance théorique des principes justifiant les solutions admises, bien que cette connaissance ne soit pas développée dans les textes qui en constituent l'application, et qui ont été découverts pour partie dès 1854 par Senkereh, pour partie en 1894.

La numération babylonienne est doublement originale : c'est une *numération de position*, et sa *base est sexagésimale*. La notation positionnelle fait dépendre la valeur d'un signe numérique de la *place* qu'il occupe dans le nombre et se différencie ainsi de la notation juxtapositionnelle, généralement répandue dans l'Antiquité, en particulier chez les Romains. Par exemple, dans 444, le 4 signifie selon sa place aussi bien 4 unités que 4 dizaines ou 4 centaines. Ce système permet l'expression aisée de nombres très grands ou très petits et facilite les opérations. Cette numération sexagésimale repose sur deux symboles particuliers :

= 1 = 10

mais l'unité peut aussi bien signifier 1 que 60, ce dont seul le contexte permet de décider (*tableau* I).

De même peut aussi bien désigner 10 · 60. Autrement dit, le besoin de disposer d'une notation pour le zéro se fait sentir, mais cette lacune n'est pas comblée.

Il semble que ce soient les Sumériens qui, après un essai de notation ternaire, aient inventé ce système sexagésimal qu'on retrouve dans leur métrologie et dont l'origine pourrait apparaître dans l'échange, l'orge, puis certains métaux servant de base aux transactions. Nous avons conservé aujourd'hui ce système dans notre partition de l'heure en soixante minutes, de la minute en soixante secondes, et dans notre mesure en degrés des angles. A cet égard, l'adjonction du zéro achève ce système et le rend propre aux applications astronomiques auxquelles procédait Ptolémée.

fig. 6

Exemple d'addition **signifie 14**

fig. 7

Symbole de soustraction

Exemple de soustraction

10 – 4 s'écrit :

fig. 8

Symbole de multiplication

Exemple de multiplication

7 fois 1 **est noté :**

Richard Colin

Les opérations arithmétiques

Les fractions ne jouent pas ici le rôle qu'elles avaient dans l'arithmétique égyptienne, la base sexagésimale permettant de se limiter souvent à des nombres entiers; mais elles ne sont pas totalement absentes, et l'on trouve même dans deux tablettes une tentative de noter les fractions par « numérateur » et « dénominateur ».

L'addition est exprimée par la juxtaposition des symboles (*fig.* 6). Mais la soustraction est signifiée par un symbole particulier (*fig.* 7).

La multiplication possède également un signe assez complexe (*fig.* 8).

La multiplication utilise d'abord des *tables* qui contiennent les produits par un nombre, *a*, des vingt premiers nombres et de 30, 40 et 50, ce qui suffit à donner le résultat du produit par *a* d'un nombre quelconque compris entre 1 et 60. Mais, pour multiplier par 37, par exemple, on multiplie par 30, puis par 7, et l'on additionne les résultats.

Pour diviser *b* par *a*, on cherche d'abord $1/a$, puis on multiplie cet inverse par *b*. C'est pourquoi aux tables de multiplication sont souvent associées des tables pour les inverses.

Les documents nous ont légué encore des tables de carrés, de racines carrées, de cubes et de racines cubiques, indiquant, s'il y a lieu, la valeur exacte d'une racine, ou, à défaut, une approximation. On a ainsi une excellente approximation de $\sqrt{2} = 1{,}414\,213...$ Les tablettes contiennent aussi des séries arithmétiques et géométriques finies, des relations exponentielles et logarithmiques. La somme d'une série géométrique de raison 2 et ayant 10 termes est calculée ainsi :

$$1 + 2 + 4 + ... + 2^9 = 2^9 + (2^9 - 1) = 2^{10} - 1,$$

résultat qui est celui de la formule connue $S = a\,\dfrac{q^n - 1}{q - 1}$, pour $q = 2$, $a = 1$ et $n = 10$.

Naturellement, nous avons renoncé, pour des raisons de place, à exposer l'histoire interne de la numération et de l'arithmétique telles qu'elles se sont développées dans les différentes civilisations mésopotamiennes, et à souligner les efforts locaux qui en ont marqué les différentes étapes.

L'algèbre

Il s'agit essentiellement de l'apparition, sinon de la théorie, des équations linéaires, quadratiques, voire biquadratiques et cubiques, du moins de ces équations elles-mêmes sous une forme si scientifique qu'elle semble l'application directe d'une théorie. La préoccupation pratique semble bornée à un rôle somme toute secondaire : elle est soit le support d'un exercice ou d'une application, soit l'auxiliaire gardant le privilège de l'illustration et de l'exemple. Lorsque, sur une tablette, on demande de trouver le côté d'un carré dont la surface « additionnée six fois et le côté trois fois et demie » vaut 906, l'indication qu'il s'agit de « mon champ » montre qu'il s'agit, non pas d'un problème pratique, mais d'un exercice théorique sur un thème mathématique donné.



Giraudon



Giraudon

▲ *A gauche, une tablette cunéiforme, vestige de la civilisation mésopotamienne (Bagdad, Musée irakien). A droite, tablette cunéiforme, d'origine babylonienne, contenant un texte mathématique et un dessin géométrique (Bagdad, Musée irakien).*

Nous donnerons un exemple de résolution d'une équation du second degré (problème 7, tablette BM 13901), tel qu'il est exposé par Labat et Bruins dans *la Science antique et médiévale* (P. U. F., 1966, p. 113) : « J'ai ajouté sept fois le côté de mon carré et onze fois la surface : cela fait 6,15. Pose 7 et 11. » Ce qui veut dire : posons

$$11\,x^2 + 7\,x = 6{,}15.$$

Le texte babylonien prescrit les démarches suivantes : « Multiplie 11 par 6,15 : 1 · 8,45 (1 · 8,45 = 68,75 en notation décimale). Prends la moitié de 7, 3,30 (ou 3,50 en notation décimale). Multiplie 3,30 par lui-même : 12,15 (= 12,25 en notation décimale). Ajoute 12,15 à 18,45 : 1 · 21 (= 81 en notation décimale). La racine de 1 · 21 est 9. Ote 0,30 que tu as multiplié de 9 : 5,30. L'inverse de 11 n'est pas dans les tables. Par quoi faut-il multiplier 11 pour avoir 5,30 ? par 0,30 (= 0,50 en notation décimale). 0,30 est le côté de mon carré. »

Cela revient à considérer l'équation de la forme

$$ax^2 + bx + c' = 0$$

suivante : $11\,x^2 + 7\,x - 6{,}15 = 0$, et les opérations suivantes :

$$11 \times 6{,}15 = 1 \cdot 8{,}45 \quad \text{qui correspond à } a \cdot c'$$

$$7{,}2 = 3{,}30 \quad \text{qui correspond à } \frac{b}{2}$$

$$\frac{7}{2} \cdot \frac{7}{2} = 12{,}15 \quad \text{qui correspond à } \left(\frac{b}{2}\right)^2$$

$$12{,}15 + 18{,}45 = 1 \cdot 21 \quad \text{qui correspond à } \frac{b^2 + 4\,ac'}{4}$$

$$1{,}21 = 9 \quad \text{qui correspond à } \frac{\sqrt{b^2 + 4\,ac'}}{2}$$

$$9 - 3{,}30 = 5{,}30 \quad \text{qui correspond à } \frac{\sqrt{b^2 + 4\,ac'}}{2} - \frac{7}{2}$$

Les Babyloniens avaient également une certaine idée de la notion de fonction, qui est généralement et à juste titre considérée comme la notion la plus caractéristique des mathématiques modernes. Pour la résolution de l'équation cubique $x^3 + x^2 = 0$, une tablette donne ainsi la liste des nombres correspondants, pour chaque valeur de n, à $n^3 + n^2$.

La géométrie

L'exposé en est également fait par schémas de solutions plutôt que par voie démonstrative au sens où nous l'entendons depuis les Grecs. Elle est dominée par une tendance algébriste, dans la mesure où ne sont mobilisés que les résultats susceptibles de conduire à des relations métriques ; les notions d'angle et de parallèle, une relation comme $a/b = c/d$ n'ont pas de termes pour les exprimer, tandis que la relation « de Pythagore » qui exprime un rapport métrique entre les côtés d'un triangle et sa diagonale est déjà connue (mais c'est Pythagore qui la démontrera). D'où la conclusion de Bruins : « A mon avis, il ne faut pas attribuer aux mathématiciens babyloniens la connaissance de la théorie géométrique des proportions, bien qu'ils montrent des relations numériques qui correspondent aux théorèmes de cette théorie. »

Parmi les problèmes où est impliquée la relation de Pythagore, citons un exemple. On donne les relations :

$$L + l + d = 40 \quad \text{et} \quad L \cdot l = 2 \cdot 0$$

(= 120 en système décimal) entre la longueur L, la largeur l et la diagonale d d'un rectangle ; le scribe résout le système des 2 équations sans indiquer qu'il suppose également donnée : $d^2 = L^2 + l^2$.

Lorsque l'application de la relation de Pythagore n'est pas possible, la valeur de la diagonale étant irrationnelle, les scribes tentent de trouver une approximation, en utilisant la formule :

$$\sqrt{x} = \sqrt{a^2 + 2} \cong a + \frac{2}{2\,a} = \frac{1}{2}\left(a + \frac{x}{a}\right)$$

On trouve évidemment dans les tablettes des problèmes portant sur des surfaces telles que des carrés, des rectangles, des trapèzes, etc., et sur des volumes tels que des cubes, des parallélépipèdes et des polyèdres. Il s'agit toujours d'une théorie essentiellement numérique. Ainsi un calcul, appliquant la formule

$$V = h\left[\frac{(a+2)^2}{2} + \frac{1}{3}\,\frac{(a-b)^2}{2}\right]$$

donne le volume du tronc de pyramide.

De manière générale, l'évaluation des surfaces n'est pas très scrupuleuse, sauf dans les formules concernant des carrés, des rectangles et des triangles rectangles. Bref, les Babyloniens ont connu une théorie numérique

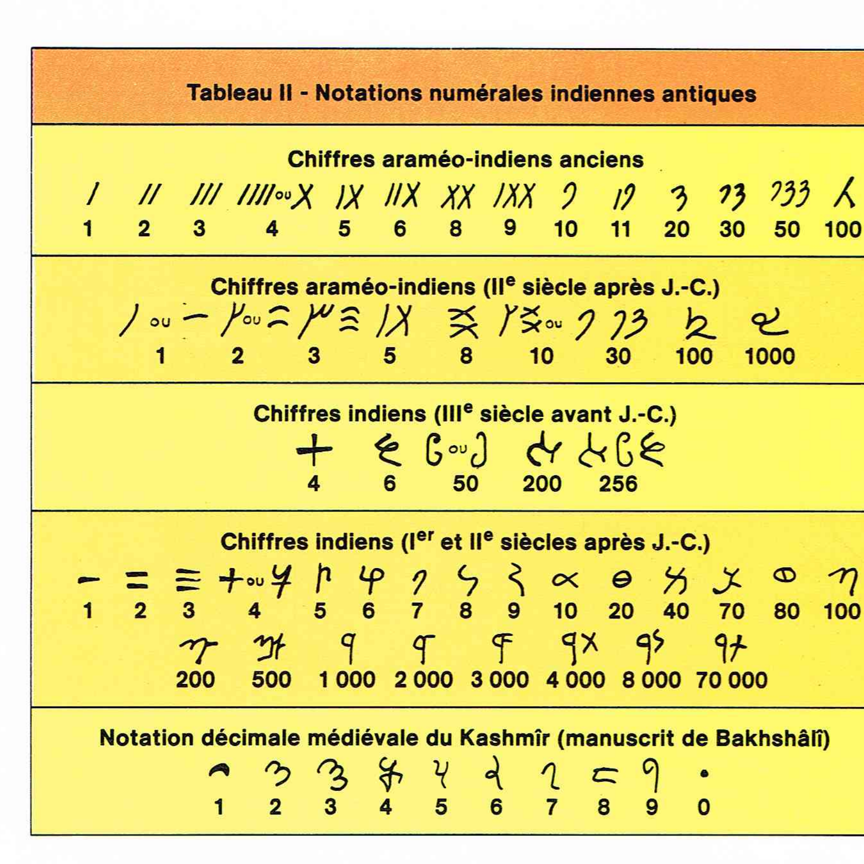


Tableau II - Notations numérales indiennes antiques

Chiffres araméo-indiens anciens

1 2 3 4 5 6 8 9 10 11 20 30 50 100

Chiffres araméo-indiens (IIe siècle après J.-C.)

1 2 3 5 8 10 30 100 1000

Chiffres indiens (IIIe siècle avant J.-C.)

4 6 50 200 256

Chiffres indiens (Ier et IIe siècles après J.-C.)

1 2 3 4 5 6 7 8 9 10 20 40 70 80 100

200 500 1 000 2 000 3 000 4 000 8 000 70 000

Notation décimale médiévale du Kashmîr (manuscrit de Bakhshâlî)

1 2 3 4 5 6 7 8 9 0

▲ *Tableau II : principales notations numérales indiennes antiques.*

des polygones réguliers; ils étaient familiarisés avec les concepts de centre et de rayon d'un cercle; ils n'ont pas de théorèmes sur des figures semblables, mais utilisent des méthodes qui procèdent, selon Bruins, « de l'application de la propriété de l'équidistance des droites parallèles et de l'additivité des aires ».

Par rapport à la mathématique égyptienne, une « avance » certaine : perfectionnement du calcul, apparition de procédés algébriques, usage de méthodes d'approximation et utilisation d'une géométrie à vocation essentiellement numérique, algébrique.

La mathématique indienne

En Inde, l'activité scientifique, qui a un caractère sacré, commence dès le milieu du second millénaire avant J.-C. Bien que nous ne possédions aucun traité spécial de mathématiques qui remonte aux périodes les plus anciennes (c'est-à-dire surtout védique et brahmanique), la langue védique apparaît déjà familiarisée avec le maniement des grands nombres : puissances de 10 jusqu'à 10^{23}. Les indications que nous possédons concernent la géométrie, science indispensable pour la construction des autels védiques. Les Çulva-sûtra, ou « aphorismes sur les cordeaux », sont spécialement consacrés aux règles de construction de ces autels; ils remontent à 700 avant J.-C., mais ne nous sont connus que d'après des rédactions ultérieures, qui se situent aux alentours de l'an 300 de l'ère chrétienne. Là aussi, nous avons affaire à une géométrie numérique. La relation de Pythagore y intervient souvent sous la forme : « La corde transversale d'un rectangle produit (par construction sur elle d'un carré) ce que produisent séparément la longueur et la largeur. »

La *numération décimale* et le *zéro*, propagés par les Arabes, n'apparaissent pas en Inde avant le Moyen Age. Le *tableau* II donnera une idée de l'évolution de la notation numérique indienne; on en retiendra essentiellement que c'est l'Inde qui a inventé le système de numération à 9 chiffres et zéro, qui est devenu universel.

C'est très tôt, au contraire, que la mathématique indienne connaît les approximations, trait qu'elle partage avec la mathématique babylonienne; ainsi $\sqrt{2}$ s'évalue de la façon suivante : $1 + \frac{1}{3} + \frac{1}{3 \cdot 4} - \frac{1}{3 \cdot 4 \cdot 34} = 1{,}414\ 2$

On trouve aussi : $\pi \simeq \frac{62\ 832}{20\ 000} = 3{,}141\ 6$, exprimé sous la forme : « Soixante-deux mille huit cent trente-deux est approximativement la circonférence dont le diamètre est vingt mille. »

La constitution d'une mathématique classique, impliquée dans des applications astronomiques, atteste, en algèbre, la capacité de résoudre par les fractions continues deux équations simultanées indéterminées du premier degré; mais ce résultat, consigné dans l'*Aryabhatiyam* d'Aryabhata (mathématicien et astronome vivant vers 476 après J.-C.), est largement dépassé par l'astronome du VIIe siècle Brahmagupta, qui parvient à une méthode pour trouver les solutions entières d'une équation indéterminée du second degré.

La mathématique chinoise

Les textes de la Chine ancienne qui nous ont été conservés, gravés sur des os, sont souvent d'inspiration divinatoire. Des nombres y sont notés, mais on ne trouve pas de liste exhaustive pour les chiffres utilisés, non plus que l'indication d'opérations sur ces nombres. Certes, il s'agit de documents anciens, remontant jusqu'à 1400 avant J.-C. Leur étude permet cependant d'établir le *tableau* III qui donne les nombres de 1 à 10.

Mais on ne possède pas pour les époques anciennes de notation systématique, et nous ne savons pas comment étaient notés effectivement des nombres comme 16, 17, 18 ou 19. En revanche, on a une notation pour les dizaines et pour les centaines.

Il s'agit donc d'une numération qui ne groupe pas toujours unités simples et puissances de la base 10, et qui ne s'est pas bien libérée de la numération parlée, qu'elle ne fait souvent que transcrire.

Si l'usage des nombres en Chine remonte à la préhistoire, des opérations sur les nombres nous ne savons pas grand-chose avant le IIIe siècle de l'ère chrétienne. Addition et soustraction se font alors directement grâce à une représentation des nombres par de petits bâtonnets ou *jonchets* qui remplissent le même rôle que les *calculi* (petits cailloux) latins; on les ajoute ou retranche colonne par colonne.

L'utilisation des jonchets sur une table permettait de figurer un nombre, rapporté comme on l'a indiqué à une base décimale, par une numération de position analogue à celle que nous employons actuellement avec nos chiffres arabes (*tableau* IV).

Pour multiplier, on posait le multiplicateur en haut et le multiplicande en bas de la tablette; les produits partiels posés sur la ligne intermédiaire étaient additionnés au fur et à mesure de leur obtention. De façon analogue, la division était conduite en posant le diviseur tout en bas, le dividende sur la ligne moyenne, et le quotient en haut.

La géométrie ancienne chinoise ne nous est connue que par les écrits de l'école de Mo Ti, qui définissent le point et la droite. Ailleurs, on ne se préoccupe que du calcul des

Tableau III - Système de numération chinoise ancienne

Ecriture courante en chiffres arabes	Nombres cardinaux		
	Ecriture	Prononciation	
		ancienne	moderne
1	一	?iět	yi
2	二	ni'	eul
3	三	sâm	san
4	四	si'	sseu
5	五	'ngo	wou
6	六	liuk	liu
7	七	ts'iět	ts'i
8	八	pat	pa
9	九	'kiǒu	kieou
10	十	žiǒp	che

▶ *Tableau III : la numération chinoise de 1 à 10 établie à partir de documents datant de 1400 avant J.-C.*

surfaces. Toutefois, des exemples simples de nombres pythagoriques (3, 4, 5, la somme des carrés de 3 et de 4 étant égale au carré de 5) apparaissent, de même qu'une grossière évaluation de π ($= 3$).

A l'époque des Han (de 202 avant J.-C. à 220 de l'ère chrétienne), apparaissent des traités de calcul, avec une arithmétique concrète des nombres rationnels et irrationnels approchés, des équations linéaires à coefficients numériques, et le plus remarquable des systèmes de n équations linéaires à n inconnues, constituées en calcul algébrique sur l'échiquier, et dont voici un exemple pour le cas $n = 3$:

$$\begin{aligned} x + 2y + 3z &= 26 \\ 2x + 3y + z &= 34 \\ 3x + 2y + z &= 39 \end{aligned}$$

Cet exemple est représenté matriciellement par la *figure* 9.

La résolution était donc faite par manipulation de jonchets, avec des jonchets de couleur pour les nombres positifs (« corrects ») et des jonchets noirs pour les nombres négatifs (« trompeurs »).

La mathématique maya

La civilisation maya a connu un système de numération positionnelle, de base 20. Les unités de 1 à 4 étaient notées par des points, ainsi que les vingtaines, et certains autres nombres comme 400 (20 × 20), 8 000 (400 × 20), etc. ; cinq et ses multiples étaient notés par des barres. Les nombres 560 (28 × 20) et 7 200 (360 × 20) sont remarquables parce qu'ils sont au principe de la partition de l'année maya en 18 mois de 20 jours, auxquels on ajoutait un complément annuel de 5 jours.

Conclusion

Le caractère général de toutes les mathématiques que nous venons de considérer rapidement, chacune replacée dans son aire culturelle spécifique, est double : il s'agit d'une mathématique toujours à la fois *utilitaire* et *sacrée*. La géométrie des Égyptiens est ainsi née entre les mains des harpédonaptes, de ceux qui « attachent le cordeau », c'est-à-dire les arpenteurs.

Mais ces arpenteurs étaient surtout préoccupés par le problème de l'*orientation* des temples, en relation étroite avec la religion égyptienne. L'arpentage lui-même est, en fin de compte, l'œuvre des prêtres; pratique rituelle et pratique manuelle ou concrète sont intimement associées. De manière générale, dans chacune de ces civilisations anciennes, la pensée mathématique était conçue, et c'est peut-être tout à leur honneur, de façon à s'inscrire strictement dans l'horizon propre de leur culture, de leur vision du monde, de l'idée qu'elles se font des rapports de l'homme et du monde, avec tout ce que cela comporte d'imbrications mutuelles entre le rite, la pratique et cet outil qu'est le calcul.

Tableau IV

Ecriture courante en chiffres arabes	Chiffres en jonchets	Troncs célestes (ordinaux)		
		Ecriture	Prononciation	
			ancienne	moderne
1	I	甲	kap	kia
2	II	乙	?iět	yi
3	III	丙	püAng	ping
4	IIII	丁	tieng	ting
5	X	戊	mŏu'	meou
6	𝍥	己	'ki	ki
7	𝍦	庚	kɛng	keng
8	𝍧	辛	siěn	sin
9	𝍨	壬	'ñiěn	jen
10	—	癸	'kwi	kouei

◀ ***Tableau IV : la représentation des nombres se faisait au moyen de petits bâtonnets ou jonchets.***

La mathématique grecque

Il est admis qu'avec la mathématique grecque, c'est l'idée d'une science autonome qui s'affirme, une science vivant de sa vie propre, *dégagée des techniques et des rites,* capable de se définir, de se distinguer, de se concevoir, au sens propre d'une *théorie*.

Mais cette mathématique ne s'est pas constituée telle quelle d'un seul coup.

Les commencements de la mathématique

Ils remontent à la fondation, généralement attribuée à Thalès, de l'**école de Milet.** Thalès aurait vécu à la fin du VIIe siècle et au début du VIe, environ de 639 à 546 av. J.-C. On raconte qu'il avait prédit une éclipse solaire, indiqué, le premier, la cause des éclipses, esquissé une première explication rationnelle de l'Univers en le réduisant à un élément unique : l'eau, génératrice des autres éléments. Si l'on en croit la tradition, Thalès aurait rapporté les premières connaissances mathématiques de ses voyages en Égypte. Proclus rapporte que « Thalès, le premier, ayant été en Égypte, en rapporta cette théorie [la géométrie] dans l'Hellade; lui-même fit plusieurs découvertes et mit ses successeurs sur la voie de plusieurs autres, par ses tentatives, tantôt plus générales, tantôt plus restreintes au concret ».

Une autre tradition, qui remonte à Eudème mais est souvent reprise après lui, lui attribue la solution des problèmes géodésiques suivants : calcul de la distance d'un navire à la côte et calcul de la hauteur d'une pyramide d'après l'ombre portée.

Ce qui est certain, c'est qu'à l'époque de Thalès, on devait disposer du *concept d'angle,* élaboré probablement en Grèce. L'attribution à Thalès des propositions d'après lesquelles le diamètre partage le cercle en deux parties égales, les angles à la base d'un triangle isocèle sont égaux, les angles opposés par le sommet égaux et l'angle inscrit dans le demi-cercle droit, est très plausible. Ces propositions correspondent respectivement à la définition 17 et aux propositions 5, 15 et 31 du premier livre d'Euclide.

Ainsi, Thalès s'est occupé avant tout des propriétés des angles, et la rupture qu'il a opérée avec la tradition égyptienne consiste essentiellement en ce qu'il a créé ce qu'on pourrait appeler une *géométrie de la ligne,* par opposition à la géométrie égyptienne, si absorbée par le calcul des aires. « Thalès, écrit Diogène, développe tout ce qui touche à la considération des lignes », ou, comme dira Apulée, des « petites lignes », c'est-à-dire que Thalès développe une étude géométrique sur figures, et non plus sur le terrain, centrée sur le *concept de similitude,* dont Bruins nous dit qu'on peut vainement le chercher dans une mathématique aussi développée que celle des Babyloniens.

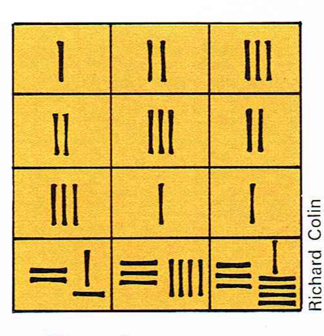
Quoi qu'il en soit, le nom de Thalès reste attaché à la première mathématique qui avait pris son essor sous l'égide philosophique de la fameuse école de Milet.

Celle-ci fut suivie par l'**école des Ioniens,** aux destinées de laquelle présidait Pythagore. D'après Proclus, qui s'autorisait du témoignage d'Eudème, c'est Pythagore qui transforma la géométrie par un « enseignement libéral », et rechercha les théorèmes abstraitement et par l'intelligence pure. Pythagore est resté célèbre par sa doctrine qui enseignait que « toutes choses sont nombres », principe en lequel se résume le pythagorisme, et qui traduit toute son ambition : exprimer en formules arithmétiques tous les phénomènes. Nonobstant les aberrantes exagérations auxquelles elle a donné lieu, cette ambition valut à l'arithmétique pythagoricienne de s'élever au rang privilégié d'une science à *caractère spéculatif.* En effet, avant de dire que les choses s'expriment par des nombres, le pythagorisme avait commencé par considérer les nombres eux-mêmes comme des réalités tangibles et intelligibles. Nous verrons comment les pythagoriciens figuraient les nombres, après que nous aurons dit un mot de la numération grecque.

La numération grecque

La numération grecque a procédé de deux systèmes.

- Le *système hérodien,* attesté dès 554 avant J.-C., ressemble au système romain que nous connaissons et présente les mêmes inconvénients que lui : il ne se prête pas à l'expression d'une arithmétique un peu avancée.



▲ ***Figure 9 : représentation matricielle d'un système de 3 équations à 3 inconnues, dont la résolution se faisait par manipulation de jonchets.***



Snark International

▲ ***L'école de Pythagore, à Croton, d'après une gravure du XIXe siècle.***

1 = I	10 = Δ	50 = 𐅄
2 = II	100 = H	5 000 = 𐅇
3 = III	1 000 = X	
5 = Π	10 000 = M	

● Le *système milésien,* né vers le milieu du Ve siècle, est plus propre à l'usage savant. Il emploie l'alphabet grec, augmenté, pour les besoins de la cause, de trois lettres auxiliaires anciennes : ϛ ϟ et ϡ.

1 = $\overline{\alpha}$	10 = $\overline{\iota}$	100 = $\overline{\rho}$
2 = $\overline{\beta}$	20 = $\overline{\varkappa}$	200 = $\overline{\sigma}$
3 = $\overline{\gamma}$	30 = $\overline{\lambda}$	300 = $\overline{\tau}$
4 = $\overline{\delta}$	40 = $\overline{\mu}$	400 = $\overline{\upsilon}$
5 = $\overline{\varepsilon}$	50 = $\overline{\nu}$	500 = $\overline{\varphi}$
6 = ϛ	60 = $\overline{\xi}$	600 = $\overline{\chi}$
7 = $\overline{\zeta}$	70 = $\overline{o}$	700 = $\overline{\psi}$
8 = $\overline{\eta}$	80 = $\overline{\pi}$	800 = $\overline{\omega}$
9 = $\overline{\theta}$	90 = ϟ	900 = ϡ

Ce système a subi des modifications qui lui ont permis d'exprimer de très grands nombres : dans l'*Arénaire* d'Archimède, on considère, par exemple, le nombre $10^8 \cdot 10^8$. Pourtant, ce n'est pas par l'amélioration de l'écriture numérique que les Grecs ont marqué l'arithmétique : ils ne sont parvenus ni à un système positionnel pur, ni à l'adoption systématique et fructueuse du zéro. Leur originalité est ailleurs, et s'affirme, précisément, dès Pythagore.

L'arithmo-géométrie des pythagoriciens

L'école pythagoricienne était située à Croton, en Italie du Sud (580-501 avant J.-C.). L'enseignement, réservé aux « initiés », était oral et n'a pas laissé de trace écrite. Nous connaissons l'œuvre des pythagoriciens par des écrits postérieurs, de Platon et d'Hérodote notamment.

Les découvertes attribuées à cette école convergent pour établir que c'est là qu'est née l'idée de remonter aux principes supérieurs et de démontrer abstraitement, c'est-à-dire en usant de procédés purement discursifs.

La distinction entre nombres pairs et impairs s'effectue alors, ainsi que le classement des nombres d'après des propriétés *arithmo-géométriques,* c'est-à-dire d'après une architecture discontinue des unités-points. La reconnaissance des *nombres parfaits* manifeste ainsi la structure interne des nombres par l'observation des propriétés des figures qui les représentent, et s'accompagne de l'indication explicite de procédés opératoires permettant de composer les nombres entiers selon une *loi.* Les nombres, points ou cailloux sur le sable, sont ainsi *classés* selon l'arrangement des points ou cailloux. On distingue ainsi plusieurs types.

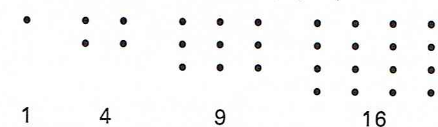
● Les *nombres triangulaires,* qui sont les exemples les plus simples de séries arithmétiques finies. 1, 3, 6, 10, par exemple, sont des nombres triangulaires :

1 3 6 10

On connaît les cas particuliers simples de la loi que nous exprimons aujourd'hui par :

$$1 + 2 + 3 + ... + n = \frac{n\ (n + 1)}{2}$$

● Les *nombres carrés,* tels 1, 4, 9, 16, etc.



1 4 9 16

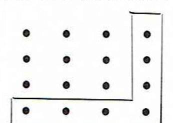
Les carrés sont des sommes de nombres impairs, par exemple : $9 = 1 + 3 + 5$; $16 = 1 + 3 + 5 + 7$; bref, on connaît des cas particuliers de la loi :

$$1 + 3 + 5 + ... + (2\,n - 1) = n^2$$

On voit aussi que tout nombre impair est l'expression de la différence entre deux carrés ayant respectivement pour côtés deux entiers consécutifs :

$$5 = 3^2 - 2^2, \quad 7 = 4^2 - 3^2, \text{ etc.}$$

Le passage d'un carré au suivant s'effectuait concrètement chez les pythagoriciens à l'aide du *gnomon* (dont le terme est emprunté aux Babyloniens), et que l'on peut représenter par la figure formée par les points supplémentaires sur le schéma ci-dessous :

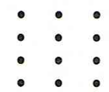


● Les *nombres hétéromèques* ou *rectangulaires :* deux nombres sont dits hétéromèques quand leur différence est égale à 1.

On sait que la somme des nombres pairs consécutifs est le produit de deux nombres hétéromèques, c'est-à-dire qu'on connaît des expressions particulières de la loi :

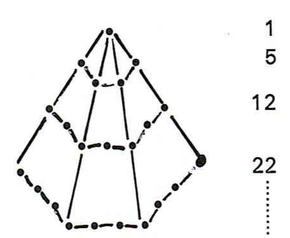
$$2 + 4 + 6 + ... + 2\,n = n\ (n - 1) ;$$

pour $n = 3$, on a la représentation suivante :

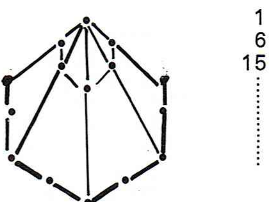


● Les *nombres polygonaux :* pentagonaux, hexagonaux, etc.

Nombres pentagonaux : 1, 5, 12, 22, ... le n-ième étant $\frac{3\,n^2 - n}{2}$:



Nombres hexagonaux : 1, 6, 15, 28, ... le n-ième étant $2n^2 - n$:



● Les *nombres parfaits* sont ceux qui sont égaux à la somme de leurs diviseurs ; par exemple :

$$6 = 1 + 2 + 3 ; \quad 28 = 1 + 2 + 4 + 7 + 14$$

Les nombres composés (= non premiers) et non parfaits sont dits *oblongs*.

● Le *concept de moyenne harmonique* qu'Archytas définissait comme « la *médiété* (on appelle médiété une progression de trois termes tels que deux d'entre eux et deux de leurs différences soient dans le même rapport) dont les trois termes sont tels que, quelle que soit la partie de lui-même dont le premier dépasse le second, le second surpasse le troisième de la même partie de ce troisième ». Autrement dit, si la différence du moyen au petit terme est une fraction donnée du petit terme, il faut que la différence du moyen au grand soit égale à la même fraction du grand terme ; par exemple, 6, 4 et 3 forment une médiété harmonique, puisque 4 surpasse 3 du tiers de 3 et est surpassé par 6 du tiers de 6.

Plus généralement, pour trois nombres a, b, c tels que $a > b > c$, on a une médiété harmonique si :

$$\frac{1}{c} - \frac{1}{b} = \frac{1}{b} - \frac{1}{a}$$

L'origine de la dénomination « harmonique » est musicale. C'est un témoin des anciennes expériences pythagoriciennes sur l'instrument de musique dénommé « monocorde », qui ont conduit l'école pythagoricienne à établir une correspondance entre les nombres et les sons, c'est-à-dire à établir les lois numériques de la musique. Sont également connus les concepts de moyenne arithmétique et de moyenne géométrique.

● Le *théorème de Pythagore* appartient au fonds le plus ancien de la mathématique démonstrative, et c'est le chef-d'œuvre, précisément, de l'école pythagoricienne que d'en avoir donné une démonstration, qu'il est peut-être hasardeux d'identifier avec celle d'Euclide (*Éléments,* livre I, propos. 47). Or, pour démontrer que « dans les triangles rectangles le carré du côté qui soutient l'angle droit est égal au carré des côtés comprenant l'angle droit », on se heurte au problème de la *duplication du carré,* et, par là, aux irrationnels.

● En effet, le cas particulier du triangle rectangle isocèle, moitié du carré considéré, et qui donne la solution de la duplication, fait découvrir *l'incommensurabilité de la diagonale du carré* à son côté. Sans doute les pythagoriciens ont-ils aussi été mis sur la voie par leur manipulation de triplets de nombres tels que (3, 4, 5), ou plus généralement tels que $\left(m, \frac{m^2 - 1}{2}, \frac{m^2 + 1}{2}\right)$, avec m impair, qui ont la propriété remarquable de vérifier le théorème de Pythagore : le carré du troisième nombre est égal à la somme des carrés des deux premiers. C'est un des plus beaux événements de la mathématique de toujours que le changement d'orientation induit par cette découverte : au lieu de calculer une valeur approchée de la mesure de la diagonale, on en a *démontré* l'incommensurabilité par rapport au côté. La démonstration procède, selon Aristote, par une réduction à l'absurde, en s'appuyant essentiellement sur la partition des nombres en pairs et impairs. « Ils prouvent, dit Aristote, que le diamètre du carré est incommensurable au côté, en montrant que s'il lui était commensurable, un nombre impair serait égal à un nombre pair. » Notons que cette preuve ne se trouve que dans les *versions tardives* du texte d'Euclide (livre X, propos. 117), et soulignons aussi que cette preuve indirecte, ou apagogique, qui n'est pas l'unique façon de démontrer l'irrationalité de la racine carrée de 2 (mesure de la diagonale du carré de côté 1), constitue un type de raisonnement employé pour la première fois par les philosophes éléates, Zénon et Parménide. (Sur les rapports de la philosophie éléatique et des mathématiques grecques, voir *Mathématiques et philosophie.*)

La géométrie des pythagoriciens est moins riche que leur arithmétique ; ils ont élaboré un segment de la théorie de la similitude, et on leur attribue également le théorème sur la somme des angles d'un triangle et certaines propositions sur les polygones réguliers (livres II et VI d'Euclide).

Les trois grands problèmes

L'école pythagoricienne fut suivie par d'autres écoles : en particulier celle des sophistes, première école athénienne, l'Académie de Platon (V[e] siècle avant J.-C.) et celle d'Eudoxe.

Au V[e] siècle, la mathématique grecque est d'abord dominée par ce qu'on a coutume d'appeler les *trois grands problèmes,* dont une solution satisfaisante ne pourra être donnée avant le XIX[e] siècle.

● Le premier problème est celui de la *quadrature du cercle.* La tradition veut que ce soit Anaxagore de Clazomènes qui aurait, dans la solitude de sa prison (il avait été accusé d'impiété), réfléchi au rapport entre l'aire du cercle et celle du carré, inscrit ou circonscrit. Après lui, Antiphon le Sophiste considère des polygones inscrits dans le cercle, Brison, des polygones inscrits et circonscrits, si bien que leur tentative de résoudre le problème de la quadrature du cercle aboutit à assimiler le cercle à un polygone régulier ayant un nombre infini de côtés : c'est la première porte ouverte aux considérations infinitésimales. Mais le problème ne se précise qu'avec les grands maîtres : Hippocrate de Chios, Eudoxe et Archimède.

● Le deuxième problème est celui de la *duplication du cube,* où l'on demande de construire un cube dont le volume est double de celui d'un volume donné, ou plus généralement un cube de même volume qu'un parallélépipède donné. L'oracle de Délos, dit la légende, ayant ordonné aux habitants de doubler l'autel du dieu, ceux-ci se seraient adressés aux géomètres. Hippocrate de Chios, si connu pour son traité sur les lunules, fut le premier à s'apercevoir que la solution revenait à trouver deux moyennes proportionnelles, en proportion continue entre deux lignes droites, dont la plus grande est le double de la plus petite, « en sorte que l'embarras fut changé en un autre » ! Autrement dit, à la recherche d'une grandeur inconnue x telle que $x^3 = abc$ (longueur, largeur et hauteur du parallélépipède), Hippocrate substitue la recherche de deux grandeurs inconnues x et y telles que :

$$\frac{a}{x} = \frac{x}{y} = \frac{y}{b}$$

Archytas donne une solution géométrique (dans l'espace à 3 dimensions) de ce problème et Ménechme, disciple d'Eudoxe, le résout par la méthode des intersections de coniques, en utilisant les paraboles :

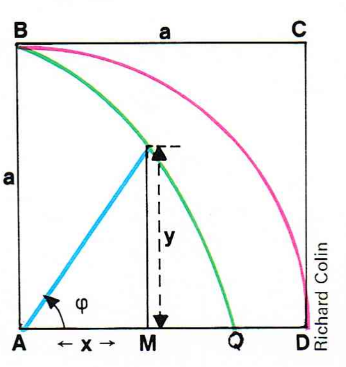
$$x^2 = ay, \; y^2 = bx,$$

et l'hyperbole $xy = ab$.

● La *trisection de l'angle* constitue le dernier de ces problèmes classiques. La trisection, à la règle et au compas, de l'angle droit ne fait pas difficulté (*Euclide,* I, 1) ; celle d'un angle obtus, décomposable en un droit et un aigu, est possible à condition de savoir faire celle d'un angle aigu ; et c'est là qu'est le problème. Hippias d'Élis en donna une solution, en inventant la *quadratrice,* qui est une courbe transcendante, non constructible à la règle et au compas (*fig.* 10).

Soit le carré ABCD, de côté a, BD l'arc de cercle de rayon a. Si AD, animé d'un mouvement uniforme autour de A, balaie l'angle droit $\widehat{DAB}$, alors que DC, en restant parallèle à sa direction, se déplace uniformément vers AB, une courbe DQ est engendrée, dont le point M est tel que :

▼ ***Figure 10 : représentation schématique du problème de la trisection de l'angle.***





Snark International

▲ ***Euclide présentant ses* Éléments *à Ptolémée Ier, d'après une gravure du XIXe siècle.***

$$\text{cotg}\ \varphi = \frac{x}{y}.$$

Comme $\frac{y}{a} = \varphi \Big/ \frac{\pi}{2}$ $\quad \frac{\varphi_1}{\varphi_2} = \frac{y_1}{y_2}$. Si bien qu'à la division de l'ordonnée $y =$ PM en n parties égales correspond celle de l'angle φ en autant de parties égales. [L'équation de cette transcendante est $x = y \cdot \text{cotg} \frac{(\pi\, y)}{2\, a}$].

Outre ces fameux problèmes, les Grecs ont créé l'étude des *sections coniques*. Ménechme les avait utilisées pour le problème de Délos ; mais c'est Éristée (350 avant J.-C.) qui sut engendrer l'ellipse, l'hyperbole et la parabole, et qui a inspiré directement Euclide. Apollonios de Perga (fin du IIIe siècle) a introduit les termes d'hyperbole, de parabole, et d'ellipse, tandis qu'Aristée et Archimède parlent de section du cône à angle aigu (ellipse), section du cône à angle droit (parabole), et section du cône à angle obtus (hyperbole). Aristée a écrit un ouvrage sur les coniques dont Pappus nous a gardé la mémoire, tandis qu'un traité d'Euclide sur le même sujet est perdu ; mais les travaux d'Archimède et d'Apollonios (8 livres sur les *Coniques*) généralisent et systématisent les résultats de leurs prédécesseurs ; en particulier, le 5e livre des *Coniques* est, avec le livre V des *Éléments* d'Euclide, la lettre sur la *Méthode* et le traité *Des spirales* d'Archimède, un des principaux chefs-d'œuvre de la géométrie grecque.

Cependant, avec ces grands noms, nous abordons la période hellénistique.

La mathématique hellénistique

On n'aurait pu mesurer à sa juste valeur tout le travail qui s'est accompli durant cette première période dont nous venons de parler, s'il n'avait abouti à ce monument euclidien qui résume aussi bien l'œuvre technique de découverte que l'œuvre méthodique d'enchaînement et de démonstration entreprises de Thalès à Eudoxe.

● *Euclide*

On sait peu de chose sur sa vie qui se situerait à Alexandrie de 365 à 300 avant J.-C. Le néo-platonicien Proclus, qui a écrit un important commentaire des *Éléments*, nous rapporte qu'Euclide avait répondu à Ptolémée Ier, qui lui avait demandé s'il n'y avait pas vers la géométrie de route plus courte que celle des *Éléments* : « Il n'y a pas de voie royale en géométrie. » Proclus nous assure aussi qu'Euclide avait ordonné et rendu cohérente une matière mathématique prête depuis Eudoxe, en complétant, entre autres, les trouvailles de Théétète et en donnant des preuves inattaquables pour nombre de propositions sans rigueur avant lui.

Les *Éléments* achèvent donc le travail fondateur des premiers mathématiciens grecs, dont les noms ne sont jamais rappelés dans cet édifice à caractère plus théorique qu'historique, au style rigoureusement démonstratif, à la portée essentiellement méthodique, à l'ambition fondatrice. Toutefois, sous l'habillage logique transparaît la *variété de l'inspiration* : les *Éléments* véhiculent, en effet, des traces de l'ancien pythagorisme et trahissent les états des discussions sur les concepts des mathématiques, et sans doute aussi sur l'idée de démonstration.

Les premiers livres (I, II, III et IV), d'inspiration pythagoricienne, traitent de la géométrie du plan. Le premier, qui étudie les triangles, les droites parallèles, les aires des parallélogrammes et des triangles, contient la démonstration du fameux *théorème de Pythagore* (propos. 47 et 48) ; les trois autres contiennent respectivement des énoncés d'arithmétique algébrique appliqués à des figures géométriques, la géométrie du cercle et l'étude des polygones à 3, 4, 5, 6 et 15 côtés.

La *notion de similitude* n'apparaît qu'au livre V, qui traite de la *théorie des proportions*. Il s'agit alors d'une mathématique issue de la réforme d'Eudoxe, et donc décalée d'environ deux siècles par rapport à celle des quatre premiers livres. La fameuse définition 5 de ce livre V signifie qu'après les entiers (ἀριθμοί) et les fractions (λόγοι) les irrationnels acquièrent droit de cité numérique sous forme de grandeurs (μόγετοι). Voici, retranscrite en langage moderne, cette définition : $\frac{a}{b} = \frac{c}{d}$ si et seulement si pour tout couple (m, n) d'entiers quelconques

$$ma > nb \rightarrow mc > nd$$
$$ma = nb \rightarrow mc = nd$$
$$ma < nb \rightarrow mc < nd$$

Il faut aussi mentionner que la définition 4 énonce l'*axiome d'Eudoxe*, plus connu sous le nom d'*axiome d'Archimède*, qui postule que deux grandeurs sont toujours comparables. Ce livre V, dont le contenu n'a été vraiment assimilé dans toute sa généralité et dépassé que depuis un siècle à peine, restera un des sommets de la pensée mathématique de tous les temps.

Le livre VI est important mais plus élémentaire ; c'est une application du livre V à des problèmes géométriques ; il contient notamment le *théorème de similitude*, improprement appelé *théorème de Thalès*.

Avec le livre VII, commence la 3e partie des *Éléments*, la partie arithmétique qui se prolonge jusqu'au livre X et constitue le premier traité, le plus rigoureux jusqu'à Gauss, de théorie des nombres. Le livre VII applique la théorie des proportions aux nombres entiers, et utilise, pour la première fois, le procédé appelé « algorithme d'Euclide » pour trouver la plus grande commune mesure de deux nombres. S'y trouvent aussi des propositions sur les nombres premiers entre eux et sur le plus petit commun multiple de deux nombres.

Le livre VIII établit une théorie des puissances entières des fractions.

Le livre IX contient, à côté de propositions sur le pair et l'impair, le théorème remarquable qui établit qu'il y a une infinité de nombres premiers.

Le livre X, célèbre, est consacré à la théorie des irrationnels ; c'est partiellement un héritage de Théétète et

fig. 11

Richard Colin

◀ ***Figure 11 :*** ***les cinq polyèdres réguliers, dont traite le dernier livre des* Éléments.**

d'Eudoxe, et qui les dépasse pour ce qui est de la minutie et de la rigueur. La première proposition fonde la méthode d'exhaustion, puis, après quelques applications de l'algorithme d'Euclide, est exposée une théorie des expressions de la forme $\sqrt{\sqrt{a}+\sqrt{b}}$. Cette proposition a un rapport étroit à l'axiome de mesurabilité (ou axiome d'Archimède). La voici dans sa formulation euclidienne : « Deux grandeurs inégales étant données, si on retranche de la plus grande une partie plus grande que sa moitié, si l'on retranche du reste une partie plus grande que sa moitié, et si l'on fait toujours la même chose, il restera une certaine grandeur, qui sera plus petite que la plus petite des grandeurs proposées. » C'est-à-dire en langage moderne : pour 2 réels a et b tels que $a > b > 0$ et pour une suite a_i, $i = 1, 2, ..., n$ et $a_i \leqslant \frac{1}{2}$ pour tout i, il existe k tel que :

$$a \cdot a_i \cdot ... \cdot a_{ik} < b.$$

Les trois derniers livres sont consacrés à la stéréométrie, ou géométrie de l'espace ; on y trouve les définitions du cône, de la sphère, du cylindre, données cinématiquement, alors que les livres de géométrie plane excluaient tout recours au mouvement. On y trouve encore l'étude, par des procédés infinitésimaux, qui, selon Archimède, remontent à Eudoxe, des aires de cercles, des volumes de pyramides, cônes, cylindres ou sphères. La technique de démonstration est la méthode d'exhaustion. Enfin, le dernier livre traite des cinq polyèdres réguliers (*fig.* 11).

Avec les *Éléments* apparaît la *méthode axiomatique*. Celle-ci, longtemps attribuée à l'influence de la théorie aristotélicienne de la démonstration (*Analytiques* II), semble plutôt répondre à la nécessité de « fermer » le discours mathématique, par des stipulations qui, soit répondent aux défis des dialecticiens, soit limitent l'investigation aux seuls domaines maîtrisés et où les êtres mathématiques aient un sens précis. Ainsi l'axiome : « Le tout est plus grand que la partie », aurait pour fonction d'exclure les spéculations dialectiques sur l'infini dans la mesure ou le nombre, tandis que le fameux axiome géométrique dit des parallèles permet la considération (abstraite) d'un espace infini.

● *Archimède de Syracuse*

L'œuvre de ce mathématicien, tué en 212 par un soldat romain, lors du sac de Syracuse, se distingue, dans la mathématique ancienne, par les considérations qui la rapprochent du calcul infinitésimal des modernes. Voici, dans l'ordre chronologique, la liste des ouvrages d'Archimède qui nous sont parvenus :

— *De l'équilibre des plans* (premier livre).
— *La Quadrature de la parabole.*
— *De l'équilibre des plans* (second livre).
— *De la sphère et du cylindre* (deux livres).
— *Des spirales.*
— *Sur les conoïdes et les sphéroïdes.*
— *Sur les corps flottants* (deux livres).
— *La Mesure du cercle.*
— *L'Arénaire.*
— *De la méthode* (lettre à Ératosthène découverte au début de ce siècle).

Deux traits principaux caractérisent le génie d'Archimède : son attention à la technique dont il sait tirer un parti théorique (la statique, par exemple, lui inspire certaines découvertes géométriques), et sa conception du continu comme somme d'une infinité d'indivisibles : une aire est une somme de segments rectilignes très petits, un volume une somme de sections planes.

Dans *la Quadrature de la parabole,* il montre ainsi que le segment de parabole, encadré par deux séries de trapèzes respectivement inscrits et circonscrits, n'est ni supérieur ni inférieur aux 4/3 du triangle ABC, et donc leur est égal (*fig.* 12).

Dans *De la sphère et du cylindre,* il établit que la surface latérale d'un cône ou d'un cylindre droit doit être supérieure à celle d'une pyramide ou d'un prisme circonscrit, et s'attaque (dans le livre II) à des problèmes comme de trouver une sphère de même volume qu'un cône ou qu'un cylindre donné. Il y parvient en ramenant la question à l'insertion de deux moyennes proportionnelles entre deux longueurs données, méthode sans doute courante à l'époque, quoiqu'elle ne s'inscrive pas dans le cadre des constructions possibles à la règle et au compas.

Il faut souligner qu'Archimède fait précéder ses résultats de l'énoncé d'*axiomes,* dont le célèbre *postulat d'Archimède* ou *axiome de mesurabilité* et la définition de la droite comme le plus court chemin entre deux points.

L'axiome de mesurabilité, qui pose que, pour deux grandeurs quelconques a et b, il existe un entier positif n tel que, multipliée par ce nombre, la plus petite des grandeurs dépasse la plus grande, est utilisé dans le *Traité des conoïdes et sphéroïdes,* où apparaissent trois nouveaux corps de révolution : sphéroïde engendré par la rotation d'une ellipse autour d'un de ses axes, conoïde obtusangle engendré par la rotation d'une branche de l'hyperbole autour de son axe transversal et conoïde rectangle engendré par la rotation d'une parabole autour de son axe. C'est dans cet ouvrage que l'application de la méthode d'exhaustion semble le plus proche du concept d'intégrale définie qui demeure implicite dans la variété des problèmes géométriques résolus.

Le traité *Des spirales,* qui est l'étude d'une courbe définie cinématiquement, la spirale d'Archimède, contient une méthode remarquable pour la détermination de la tangente en un point à la courbe, qui en fait le plus ancien traité de calcul différentiel.

Dans *la Mesure du cercle,* on trouve la première approximation de π par un encadrement entre des bornes rationnelles :

$$3 \cdot \frac{10}{71} < \pi < 3 \cdot \frac{10}{70} \cong 3{,}14.$$

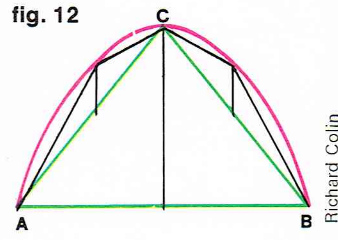
Enfin, dans *l'Arénaire,* se trouve un système de numération qui permet l'expression des très grands nombres, par exemple l'unité suivie de 800 millions de zéros, illustrée par le calcul d'un nombre supérieur à celui des grains de sable contenus dans la sphère des fixes.

● *Apollonios de Perga*

Apollonios (fin du IIIe siècle-début du IIe) est surtout connu par son fameux traité sur les *Coniques* (dont il a déjà été question ici), partiellement conservé en grec et partiellement en arabe. D'autres travaux, fort nombreux, ne nous sont connus que par le commentaire de Pappus. Apollonios a systématisé et unifié l'étude des coniques, considérées comme sections d'un cône de révolution par un plan perpendiculaire à une de ses génératrices. Le livre V est un des principaux chefs-d'œuvre de la mathématique ancienne ; on y construit la normale en un point à une courbe, non pas comme perpendiculaire à la tangente en ce point, mais indépendamment d'elle, comme la distance la plus courte d'un point donné à la courbe. Il y a là un style d'inspiration si profonde que, pour la dépasser, il n'a pas suffi, comme on le croit trop souvent, de la géométrie cartésienne ; il y a fallu encore les notions, plus tardives, de transformation et d'éléments idéaux.

● *Diophante*

Au début du IIe siècle de l'ère chrétienne, les préoccupations calculatoires connurent un nouvel essor avec Nicomaque de Gérase (vers l'an 100), le très élémentaire



▲ ***Figure 12 :*** ***le problème de la* quadrature de la parabole *fait l'objet d'un des ouvrages d'Archimède qui nous sont parvenus.***

▶ *Page extraite d'un manuscrit persan du XVe siècle : la géométrie d'Euclide commentée par At-Tusi (bibliothèque Millet, Istanbul).*

Théon de Smyrne, et surtout le grand Diophante (milieu du IIIe siècle), dont les *Arithmétiques* viennent tout juste de s'enrichir d'un manuscrit arabe contenant certains livres qu'on croyait perdus. On ne dispose pas à ce jour d'une édition scientifique de son œuvre connue. On dit que celle-ci résulte de la renaissance pythagoricienne du IIe siècle, et on y voit aussi la continuation de la tradition des logisticiens pour qui le mot « nombre » avait un sens large, englobant l'inconnue des problèmes algébriques dont la solution peut être entière, fractionnaire, voire même irrationnelle. Mais, ce faisant, on suppose sur l'histoire de l'algèbre bien des choses qui sont loin d'être établies.

Il n'y a pas de doute, néanmoins, que Diophante résout des problèmes « indéterminés » du type : $x^3 + y^3 = x + y$. Bien que l'on appelle aujourd'hui équations diophantiennes des équations aux solutions entières : $ax + by = c$, on ne trouve chez lui que des équations de la forme : $x^2 - ay^2 = 1$ (appelées *équations de Pell*). La méthode de Diophante consiste à chercher des solutions particulières, en usant d'artifices qui n'ont jamais le caractère général des vraies méthodes algébriques, et qui souffrent du manque patent d'une notation adéquate, ce qui n'avait pas empêché leur auteur de s'attaquer à des problèmes vraiment difficiles.

● *Pappus*

Après Diophante, c'est Pappus (vers 320 de l'ère chrétienne) qui achève la période vraiment créatrice de la mathématique hellénistique, et grecque en général. On trouve notamment dans la *Collection mathématique* une généralisation du théorème de Pythagore, les théorèmes dits aujourd'hui de Guldin sur la relation entre les centres de gravité et les aires ou les volumes des corps de révolution, et le théorème dit de Pappus-Pascal. Puis vient la période des commentateurs : Proclus, qui a commenté Euclide, Eutocius, qui a commenté Archimède et Apollonios.

Conclusion

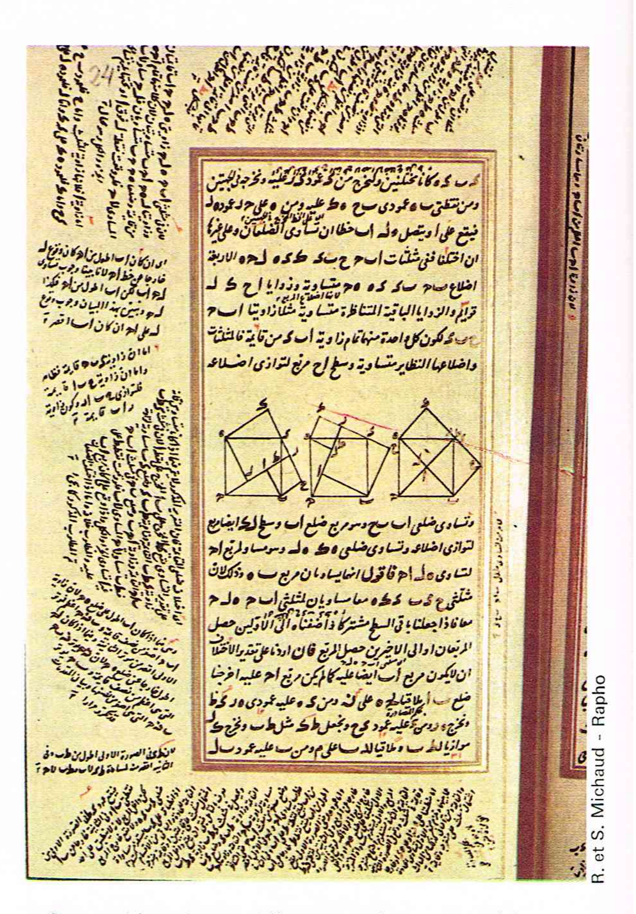
Quand on veut caractériser globalement la mathématique grecque, on dit que l'essentiel en est l'usage d'un raisonnement hypothético-déductif et des démonstrations. C'est dire qu'elle a inauguré une démarche qui, après avoir accepté un point de départ plus ou moins justifié, parcourt par ordre une série d'étapes, dont chacune doit être, de manière vérifiable, compatible avec le point de départ, et où tout nouvel apport de l'intuition est exclu. Méthode maintes et maintes fois évoquée dans les *Dialogues* de Platon, mais rarement de façon aussi précise que dans ce passage du *Phédon*, où Socrate nous dit que sa découverte essentielle consiste dans l'extension de cette méthode selon laquelle « après avoir dans chaque cas pris pour base une notion, celle qu'éventuellement je juge la plus forte, tout ce qui, selon moi, a consonance avec elle, je le pose comme étant vrai... tandis que si la consonance fait défaut, je pose que ce n'est point vrai » (*Phédon*, 100 b).

On a reproché pourtant à cette mathématique d'avoir insuffisamment élucidé les notions primitives, de ligne droite, par exemple, de surface, de rapports de grandeurs. Et on y a vu aussi une étude qualitative de la quantité, par opposition à l'esprit de la science classique, qui est une étude quantitative de la qualité.

La mathématique médiévale

La mathématique arabe

La civilisation brillante qui se développe, à la fin du VIIIe siècle et au début du IXe, sous l'impulsion unifiante de l'Islam, a fondu ensemble des courants de culture de toute origine, et d'abord des traditions mathématiques divergentes dont on n'aurait point auparavant imaginé la rencontre possible. Non seulement on rassemble avec enthousiasme l'héritage de l'Hellade, mais encore on l'enrichit du contenu de textes mathématiques néo-persans ou sanscrits. Une nuée de traducteurs donne en version arabe Euclide, Archimède, Ménélaus, Pappus, Ptolémée, si bien que l'on dispose souvent aujourd'hui de manuscrits arabes plus anciens que les textes grecs retrouvés, et que, si ceux-ci semblent irrémédiablement perdus, c'est aux traductions arabes que l'on s'adresse pour compléter ce que l'on sait de la mathématique grecque.



R. et S. Michaud - Rapho

Certes, bien des problèmes posés par ces faits sont de nature plutôt philologique. L'histoire des mathématiques arabes se réduirait à des questions de philologie si elle n'était que l'histoire de la transmission du savoir grec. Or, l'œuvre d'Al-Khawārizmi, par exemple, est si originale que les sources indiennes et grecques ne peuvent expliquer la nouveauté d'un horizon dont elles ne forment que des éléments, repensés et, pour ainsi dire, recréés.

Il est cependant trop tôt pour présenter un exposé de la mathématique arabe, dont l'étude se limite encore, la plupart du temps, à établir des listes de noms et de titres et à exploiter l'ample matière léguée par les polygraphes. Nous ne disposons pas encore d'une synthèse valable. Celle d'A. P. Juskewitsch (*Geschichte der Mathematik im Mittelalter* ou *Histoire de la mathématique au Moyen Age*, Leipzig, 1964) reste la meilleure, bien que les détails en soient souvent à revoir. Nous centrerons nos indications uniquement sur les points suivants :

— le commencement de l'algèbre ;
— les généralisations ;
— la naissance de la géométrie algébrique.

Le commencement de l'algèbre

Les débuts de la mathématique arabe sont liés au nom de Mohammed Ibn Moussa Al-Khawārizmi dont le manuel d'arithmétique (de 830 environ) nous est connu par une traduction latine du XIIe siècle. C'est le premier manuel clairement fondé sur le principe de position. Il débute par une description du système de numération dite indienne, au moyen des symboles : 1, 2, 3, 4, 5, 6, ... 9, et du « petit cercle » qui sert à noter le zéro. Suit l'exposition des opérations arithmétiques sur les nombres entiers. Notons, sans entrer dans les détails, que ce manuel a joué un très grand rôle dans le développement de l'arithmétique.

Mais Al-Khawārizmi est surtout l'auteur du premier traité d'*Algèbre*. Immortalisé une première fois par son nom même qui, défiguré, sert à désigner un des concepts les plus anciens et les plus actuels de la mathématique : le concept d'« algorithme », il l'a été aussi une deuxième fois par le titre de son second ouvrage : *Al-jabr wal muqābala*, ce qui signifie « disposition et contraposition » en entendant par le premier mot le fait qu'une équation ne contient que des termes positifs et par le second le fait de soustraire des quantités égales de chaque côté du



R. et S. Michaud - Rapho

◀ *Page extraite d'une traduction latine du XIIe siècle du* **Traité d'algèbre d'Al-Khawārizmi** *(Bibliothèque nationale, Vienne).*

signe =. La dénomination « algèbre » apparaît donc dans le titre et semble avoir été choisie pour désigner la branche des mathématiques où l'on calcule avec des lettres. L'ouvrage a exercé une grande influence sur les mathématiciens du Moyen Age. Malgré l'absence d'un symbolisme adéquat dont la fonction est remplie par les mots de la langue, Al-Khawārizmi atteint ici un degré de systématisation sans précédent; toute équation du premier ou du second degré est ramenée à l'un des 6 types suivants :

1) $ax^2 = bx$	4) $ax^2 + bx = c$
2) $ax^2 = c$	5) $ax^2 + c = bx$
3) $bx = c$	6) $bx + c = ax^2$

La réduction à ces formes canoniques se faisait par l'*opération al-jabr* qui consiste à transporter dans l'un des membres de l'équation, sous forme de termes à additionner, les termes qui sont à soustraire dans l'autre membre, et par l'*opération al-muqābala* qui consiste à retrancher aux deux membres les termes égaux. Al-Khawārizmi indique les conditions d'existence des racines : les règles sont énoncées sur des exemples mais sous une forme générale; certaines sont démontrées à l'aide de transformations géométriques. Cependant, il n'est question que des racines positives. L'ensemble tend vers la constitution d'un véritable calcul algébrique.

L'algèbre des équations du second degré fut, ensuite, développée par Abū Kāmil. Le traité de celui-ci ne contient déjà plus les applications géométriques que l'on trouvait dans celui d'Al-Khawārizmi. Mais le moment le plus important est celui où cette algèbre devient un objet mathématique propre, c'est-à-dire celui où les démonstrations ne recourent plus aux constructions géométriques; c'est alors que se dessine un second mouvement dans l'histoire de l'algèbre arabe : une arithmétisation permettant d'opérer sur les inconnues par tous les moyens de l'arithmétique opérant sur les connues, selon le vœu de l'algébriste marocain As-Samaw' al.

Les généralisations

Cette généralisation est, en fait, précédée par l'œuvre du fameux Omar Khayyām, chez lequel l'idée de l'algèbre comme discipline autonome devient particulièrement manifeste. « L'algèbre, écrivait-il dans son ouvrage (1074), est un art scientifique. Son objet est : le nombre absolu et les grandeurs mesurables, étant inconnus, mais rapportés à quelque chose de connu de manière à pouvoir être déterminés; cette chose connue est une quantité ou un rapport individuellement déterminé, ainsi qu'on le reconnaît en l'examinant attentivement; ce qu'on cherche dans cet art, ce sont les relations qui joignent les données des problèmes à l'inconnue, qui de la manière susdite forme l'objet de l'algèbre. » Un objet conçu, comme on le voit, sans recours ni à la géométrie ni à l'arithmétique; l'algèbre apparaît désormais comme une théorie des équations.

C'est pourquoi l'essentiel du travail d'Omar Khayyām consiste dans une classification des équations. Comme on doit distinguer des grandeurs numériques et des grandeurs continues, l'algèbre exige une solution numérique et une construction géométrique correspondant à l'équation. Khayyām montre ainsi qu'il y a une relation entre la solution par radicaux des équations du deuxième degré et la construction géométrique; cela le conduit à s'apercevoir que la solution par radicaux des équations cubiques ne peut être trouvée, lacune qu' « un autre qui nous succédera comblera »! Il expose alors la théorie géométrique des équations du troisième degré, puis dénombre 19 types d'équations canoniques à racines positives : 5 se réduisent à des équations quadratiques ou linéaires, 6 sont des trinômes, 7 des quadrinômes, et une correspond à l'équation binomiale $x^3 = a$. Il explique, en outre (pour la première fois, notent Juskewitsch et Rosenfeld), que les équations du troisième degré ne peuvent pas être résolues à l'aide des « propriétés du cercle », c'est-à-dire par des radicaux quadratiques. Dans ce domaine, la preuve, nous dit Khayyām, ne peut être fournie que par les propriétés des sections coniques. Descartes le répétera en 1637, et en 1837 c'est Wantzel qui le montrera.

L'extension du calcul algébrique, qui fut l'œuvre d'Al-Karaji et d'As-Samaw'al, donna enfin à la science algébrique toute sa spécificité. La théorie des équations se constitua par l'application progressive des différentes opérations de l'arithmétique aux termes et expressions algébriques. C'est ce que montre l'édition par M. M. Salah Ahmed et Roshdi Rashed (Damas, 1972) d'un manuscrit intitulé *Al-Bahir en algèbre*, dû à As-Samaw' al; il en ressort que :

— Les algébristes arabes ont parfaitement défini les exposants négatifs et nuls, avant Chuquet et Stifel, et ont donc fait la première étude systématique des exposants algébriques. As-Samaw' al va plus loin que ses prédécesseurs, car il définit par induction $x^n = x^{n-1}x$, pour $n = 1, 2, 3, \ldots$; il étend la notion de puissance algébrique d'une quantité à son inverse, énonce la règle de multiplication et de division des puissances algébriques dans sa généralité, dégage explicitement l'idée d'addition algébrique des puissances, en exploitant, par la méthode des tableaux, l'isomorphisme de groupes entre $(\mathbf{Z}, +)$ et $(\{x^h ; x \in \mathbf{Z}\}, x)$.

— C'est sur cette base que peut s'édifier une théorie générale de la divisibilité des expressions algébriques, de l'approximation des fractions entières, de l'extraction de la racine des polynômes à coefficients rationnels, et de l'extension du calcul algébrique aux quantités irrationnelles algébriques. La question de savoir comment utiliser les instruments arithmétiques dans les quantités irrationnelles s'est même posée explicitement.

Naissance de la géométrie algébrique

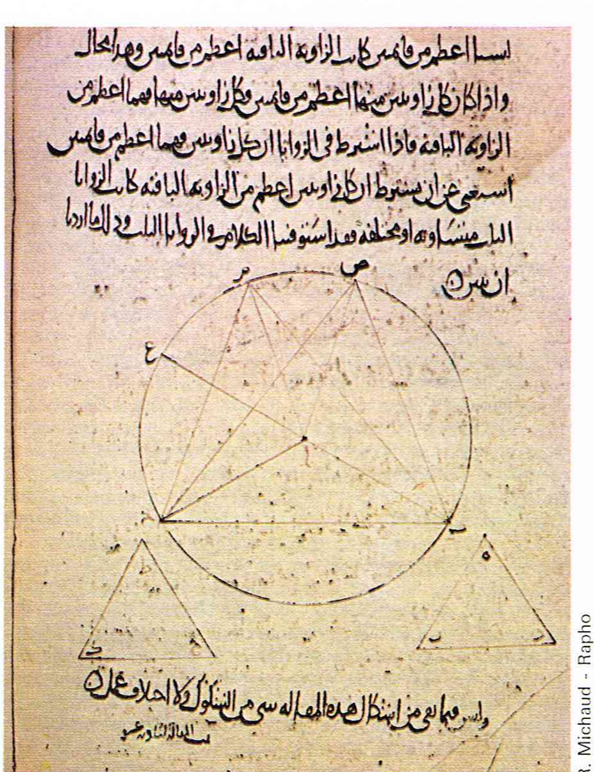
Les problèmes posés par les équations du troisième et du quatrième degré ont sans doute orienté l'intérêt vers une géométrie algébrique. Les Arabes ont ainsi retrouvé la solution d'Archimède du problème que constitue la détermination de l'intersection d'une sphère par un plan, de sorte que le rapport des volumes des deux segments sphériques obtenus égale un rapport préalablement donné. On cite généralement Al-Mahānī pour avoir ramené ce problème à la résolution d'une équation du type :

$$x^3 + 2 = px^2,$$

et Ibn Al-Haytham pour l'avoir résolu par une construction géométrique représentant la grandeur x par l'abscisse du point d'intersection de deux sections coniques convenablement choisies.

C'est dans un traité réservé à la théorie des équations que Saraf-ed-Din At-Tūsī, persévérant dans cette voie, ébauche une géométrie algébrique par l'usage d'une méthode qui s'apparente à celle de Viète. At-Tūsī applique sa méthode des tableaux à la solution de l'équation

▶ *A gauche, page extraite d'un manuscrit de géométrie arabe datant du XIIIe siècle (Bibliothèque universelle, Istanbul). A droite, page d'un livre de* Boèce *sur l'arithmétique et la géométrie, imprimé à Augsbourg en 1488.*



R. Michaud - Rapho

Snark International

$x^2 + a_1x = N$ (avec $a_1 \in Z$), et l'étend aux équations cubiques. L'exemple de la solution de l'équation : $x^3 - a_1x - a_2x - c = 0$ montre qu'en fait cette méthode s'applique à une fonction polynôme à coefficients rationnels (dans Z), et, par ailleurs, qu'elle véhicule implicitement la notion importante de dérivée. At-Tūsī, en effet, utilise systématiquement pour diviseurs des expressions qui correspondent algébriquement à la dérivée première ; dès lors, la détermination de la racine entière positive d'une équation numérique par une méthode d'approximation véhicule aussi la notion de fonction. Du reste, dans la discussion de l'*existence des racines* des équations algébriques, la notion de dérivée apparaît nettement ; At-Tūsī, en classant les équations, isole celles qui n'ont de racines que sous certaines conditions ; pour les résoudre, il est conduit à étudier la valeur maximale des expressions algébriques, à prendre leur dérivée première, à l'annuler, puis à montrer que la racine de l'équation obtenue, substituée dans l'expression algébrique, donne le maximum cherché. Tout cela le met sur la voie des notions de « limite », de borne supérieure et borne inférieure des racines d'une équation, et, par suite, sur la voie d'une étude algébrique des courbes, de la découverte de la formule de Cardan, de la corrélation entre transformation affine, divisibilité et dérivée dans la solution d'une équation. D'où l'usage des équations des courbes dans la démonstration, par exemple, de l'intersection de la parabole et de l'hyperbole.

Ainsi, c'est chez les Arabes que l'arithmétique et l'algèbre se sont émancipées de la tutelle géométrique. Et c'est dans le cadre de cette émancipation qu'est née une algébrisation de la géométrie, c'est-à-dire une préfiguration de la géométrie algébrique. On peut aussi bien ajouter que la trigonométrie ne s'est développée comme discipline autonome qu'entre les mains des Arabes : c'est Al-Khawārizmi qui donne les premières tables de sinus, qui furent traduites en latin dès 1126, et le traité de At-Tūsi (*Traité du quadrilatère complet,* 1260) a exercé une influence considérable sur le développement de la trigonométrie.

Parmi bien d'autres prouesses de calcul sur lesquelles nous ne pouvons nous arrêter ici, nous ne retiendrons qu'une remarquable approximation de

$$\pi = 3{,}141\,592\,653\,589\,793\,25,$$

un calcul original de la quadrature du segment de parabole, équivalent au calcul de l'intégrale $\int_0^a \sqrt{x}\,dx$, le calcul du volume de certains nouveaux corps de révolution obtenus par la rotation du segment de parabole limité par une corde et par le diamètre conjugué, autour de cette dernière droite, équivalent à l'intégration : $\int_0^a x^4\,dx$.

L'Occident médiéval

On ne trouvera rien d'aussi vivant dans la mathématique de l'Occident médiéval. Ou plutôt presque rien avant la constitution de la scolastique, c'est-à-dire avant la fondation des universités de Paris, de Bologne et d'Oxford. Dans le Haut Moyen Age, rares sont ceux qui, comme Boethius, se préoccupent de sauver et transmettre l'héritage ancien. La recrudescence de l'anarchie limite le développement scientifique, mais favorise l'application non scientifique des nombres aux textes sacrés, l'assimilation des distances entre les astres à une gamme musicale.

Il faut attendre les premiers contacts avec l'Islam pour retrouver des faits dignes d'attention. Gerbert (945-1003), archevêque de Reims, de Ravenne et pape en 999, sous le nom de Sylvestre II, a séjourné en Espagne de 967 à 969 ; c'est lui qui a diffusé en Europe les chiffres arabes, et il fut le premier, au témoignage de Guillaume de Malmesbury, à prendre l'abaque aux Sarrasins. De fait, le développement du commerce ne cesse de poser des problèmes, de la tenue des livres au calcul des monnaies, si bien que l'abaque lui-même cède bientôt la place à l'*algorisme* et que des simplifications des procédés de calcul se développent jusqu'à livrer l'outillage intellectuel qui permettra la révolution de pensée qui caractérise la science classique.

Après la prise de Tolède (1085) commence une nouvelle circulation du savoir sur le pourtour de la Méditerranée. Puis, au XIIe siècle, se développe le mouvement des traductions de textes arabes. Un moine, Walcher de Malvern, se fait enseigner comment les Arabes calculent le cours du Soleil et de la Lune, et conçoit, dans sa fameuse *Disciplina claricalis,* une nouvelle classification des branches du savoir qui favorise les sciences exactes, avec, aux premières places, la logique, l'arithmétique, la géométrie et la médecine, sans négliger, cependant, la musique, l'astronomie et, en dernier lieu, la philosophie ou grammaire.

D'un autre côté, Léonard de Pise (Leonardo Fibonacci, mort vers 1250) s'est initié directement, par des séjours en Afrique du Nord et au Moyen-Orient, aux pratiques du calcul. A son retour à Pise, il compose le célèbre *Liber abaci* (1202), qui familiarise l'Europe avec les chiffres

arabes ; dans ses calculs, il utilise le zéro et la suite dite de Fibonacci. Cet ouvrage, remanié et réédité en 1228, contient des connaissances mathématiques précieuses qui font de Léonard l'ancêtre de l'école algébrique italienne.

Plus tard, les préoccupations scientifiques gagnent le nord de l'Europe. A Paris, dès avant Nicolas Oresme († 1382), Johannes Campanus et Jordanus Nemorarius étudient la mathématique antique ; à Oxford, Thomas Bradwardine († 1349) s'intéresse à des notions infinitésimales dans son *De continuo*, qui rappelle les grandes discussions de Zénon et d'Aristote sur l'infini. Le trait le plus original se trouve certainement dans l'ouvrage d'Oresme, appelé *Tractatus de latitudinibus formarum*, qui présente le premier système de coordonnées, en montrant la nécessité d'exprimer les nombreuses variations des phénomènes par le recours au concept de figure géométrique. On se trouve sur la voie d'une description graphique des phénomènes naturels, donc d'une *mathématisation de la physique*. C'est le commencement d'un essor intellectuel dont l'ampleur apparaîtra dans la période suivante, période décisive qui est celle de la Renaissance et de la mathématique classique.

La Renaissance et la préparation de la mathématique classique

On insiste souvent sur le caractère révolutionnaire de l'orientation scientifique du XVIIe siècle, et le fait est qu'est en jeu une transformation intellectuelle qui rend la mathématique organiquement liée au souci d'agir efficacement sur le monde, de suivre et d'informer l'expérience, de la géométriser. Il ne faudrait cependant pas oublier que la science classique est l'aboutissement d'efforts propres à la Renaissance, laquelle prolongeait le mouvement d'appropriation du savoir arabe et grec commencé au Moyen Age.

Au réveil des études mathématiques semble présider Nicolas de Cues (1401-1463), qui naquit sur les bords de la Moselle et fut une sommité intellectuelle de son temps. Ses ouvrages strictement mathématiques n'ont pas fait époque, tandis que ses considérations métamathématiques sur la notion d'infini et sur l'idée que seules les mathématiques permettent à l'homme de comprendre l'Univers semblent avoir eu une grande influence sur des penseurs aussi importants que Léonard de Vinci, Giordano Bruno, Copernic et Kepler.

Pour noter quelques aspects de l'essor préparatoire de la Renaissance, il convient d'évoquer l'œuvre de Georg Peurbach (1423-1461), dont l'effort consista essentiellement dans l'assimilation de l'héritage gréco-arabe, dans la diffusion du calcul, et surtout dans la rationalisation de l'astronomie. Profitant des travaux arabes, surtout ceux d'Arzachel, il a écrit l'un des premiers traités de trigonométrie européens, accompagné d'une table des sinus d'une grande précision pour l'époque.

Regiomontanus (1436-1476), son brillant disciple, Johannes Müller de son vrai nom, a travaillé directement sur les écrits de Ptolémée, a édité et commenté des textes scientifiques grecs ; son œuvre personnelle est contenue dans le *De triangulis omnimodis libri quinque* (imprimé seulement en 1533) où, à la suite de Nasir ad Din At-Tusi, il veut constituer la trigonométrie comme science autonome, tout à fait indépendante de l'astronomie. A. Koyré a souligné que Regiomontanus connaissait très bien ses prédécesseurs juifs et arabes auxquels il emprunte beaucoup de ce que l'on lui a attribué, mais son œuvre eut, en tant que somme systématique, une influence méritée. Les deux premiers livres du *De triangulis* sont consacrés à la trigonométrie plane, les autres à la trigonométrie sphérique ; on y trouve le théorème fondamental, déjà connu de Levi-ben Gerson, sur la proportionnalité des côtés du triangle aux sinus des angles correspondants.



Giraudon

◀ *La « Géométrie », dans une miniature du traité de Cassiodore sur les arts libéraux (Paris, Bibliothèque nationale).*



Pedicini

▲ *Portrait de* **Luca Pacioli,** *mathématicien italien (1445-1510) dont le principal ouvrage,* **Summa de arithmetica, geometria, proportioni et proportionalita,** *résume la somme des connaissances mathématiques de son temps.*

L'œuvre de Luca Pacioli, la *Summa de arithmetica, geometria, proportioni et proportionalita* (1494), moins originale que le *Triparty* de Chuquet (1484), a cependant ce caractère de synthèse à la fois méthodique et savante qui lui assure une large diffusion et en fait une transition vers le XVIe siècle : elle contient un système d'appellations qui préfigure l'effort de codification du langage mathématique et emploie des signes pour noter certaines opérations (*p* et *m* pour + et —, par exemple).

Léonard de Vinci (1452-1519), ami de Pacioli, et, en Allemagne, Albrecht Dürer (1471-1528), sont de ces génies, artistes et amateurs de géométrie, sachant jeter un pont entre la théorie pure et les techniciens, artisans, peintres, architectes, dans le but, en définitive, d'intégrer l'art lui-même à la science. Léonard de Vinci, formé dans l'atelier de Verrocchio, atelier de peinture en même temps qu'école de tous arts et métiers, s'intéresse à une géométrie dont il aperçoit la continuité avec la mécanique. D'où la force de son intuition qui lui fait découvrir le centre de gravité de la pyramide, et la solution mécanique du problème d'optique dit problème d'Al-Hazen, dont Huygens seulement donnera la solution géométrique.

L'autre témoin de cette nouvelle sensibilité à la réalité géométrique spatiale, Albrecht Dürer, accorde un intérêt tout particulier à l'étude des figures de l'espace dans son célèbre *Underweysung der messung mit dem Zirckel und richtscheyt* (1525). Il fonde les applications esthétiques sur les règles qui permettent de construire les courbes, les surfaces et les solides utiles à cette fin ; cela le conduit à une élaboration des méthodes de perspective, en même temps qu'à une première conception des procédés de géométrie descriptive élémentaire.

Synthétique, tournée vers une vue géométrique des objets et des mécanismes, à vocation pragmatiste et de masse, la mathématique en gestation à la fin du XVe et au début du XVIe siècle se trouve favorisée par l'impression des œuvres nouvelles et des œuvres de l'héritage ancien, et surtout par la conscience du besoin d'un enseignement de mathématiques pures : une chaire de mathématiques

est créée à Bologne en 1496, puis à Vienne, à Paris, à Heidelberg, à Coimbra. Le langage mathématique est forcé d'évoluer : à côté des abréviations mathématiques déjà employées par Diophante, on recourt à de nouvelles notations; la voie de l'*algèbre symbolique* se dessine dès Chuquet, comme nous l'avons dit, et s'affirme avec Stifel, qui emploie les signes + et —, le signe $\sqrt{}$ pour la racine d'un nombre, les lettres A, B, C répétées en nombre de fois égal au degré, pour désigner les inconnues. Chuquet et Stifel mettent en parallèle progression arithmétique et progression géométrique, ce qui mènera plus tard à l'invention des logarithmes.

Le pas le plus décisif s'accomplit dans l'école italienne, avec une pléiade de mathématiciens remarquables, tels Scipio del Ferro, Tartaglia, Cardan, Ferrari, Bombelli, qui ont donné la solution de l'équation du troisième degré par le moyen des radicaux cubiques, ce qui constitue une avance incontestable mais n'a pas l'originalité absolue que les historiens lui avaient reconnue. C'est d'abord Ferro, puis Tartaglia qui redécouvrent la solution de l'équation de la forme $x^3 + ax = b$, mais sans rien publier; Cardan intervient publiquement en 1539, par sa *Practica arithmeticae generalis,* où, avec un incontestable talent d'algébriste, il montre une capacité à concrétiser les problèmes, à choisir l'inconnue, à utiliser toutes les astuces de calcul et à introduire des grandeurs auxiliaires et des changements de variable. L'*Ars magna* énumère toutes les formes possibles de ces équations du troisième degré et les traite par des méthodes numériques souvent très proches des nôtres.

Le mouvement est couronné par les travaux de Raphaël Bombelli. Ce mathématicien ingénieur admit l'existence des imaginaires et éclaira ainsi le cas irréductible des équations du troisième degré. Les règles de calcul pour ces êtres mathématiques nouveaux correspondent à celles que nous suivons aujourd'hui. Vers la fin du XVIe siècle, l'Italie ne compte plus de mathématicien d'aussi grande envergure, et la tradition s'enracine dans le Nord.

Simon Stevin (1548-1620), ingénieur militaire et inspecteur des digues dans les États de Hollande, enseigne en 1600 les mathématiques à Leyde (dans la langue flamande). Tourné vers la pratique, le concret, il publie des tables d'intérêt, traite des méthodes de comptabilité, mais révèle des talents de théoricien dans ses *Problèmes géométriques* (1583), où il perce l'analogie entre le continu et le discontinu; son *Arithmétique* occupe une place importante dans l'histoire des mathématiques, car elle réalise une systématisation et une simplification de l'arithmétique et de l'algèbre, et apporte deux innovations importantes : l'introduction systématique des fractions décimales et une nouvelle conception du nombre qui admet les incommensurables au même titre que les entiers et les fractions, car « l'incommensurance ne cause pas absurdité », et s'élève contre la qualification de nombres tels $\sqrt{2}$ ou $\sqrt{8}$, comme « absurdes, irrationnels, irréguliers, inexplicables ou sourds ». Cette unification est d'importance pour le développement de l'algèbre et de la géométrie analytique.

Ainsi se sont élaborées une pratique et des techniques de calcul que l'époque suivante n'aura qu'à examiner pour en formuler les principes implicites.

La science classique

Si tout est préparé par la Renaissance, si une essentielle conquête du réel s'y accomplit qui met les mathématiques au service d'un essor artistique sans précédent, d'une investigation de l'Univers démystifié et ouvert à la navigation transocéanique, aux échanges désormais mondiaux, et, plus généralement, au service d'une éducation au concret qui entraîne une réorganisation du temps et de l'espace, c'est seulement au XVIIe siècle qu'a lieu l'initiation à la modernité et que s'imposent les traits essentiels de la science classique.

Il faut d'abord souligner à quel point la mathématique est, désormais, imbriquée dans les préoccupations spécifiques aux sciences de la nature et combien elle est étroitement liée à la mécanique, à l'optique et à l'astronomie. Son rapport à l'expérience est indiscutable, quoique de nature complexe. Elle devient une *science utile,* porteuse du grand rêve cartésien : « parvenir à des connaissances qui soient fort utiles à la vie » et « au lieu de cette philosophie spéculative qu'on enseigne dans les écoles... en trouver une pratique, par laquelle, connaissant la force et les actions du feu, de l'eau, de l'air, des astres, des cieux et de tous les autres corps qui nous environnent, nous les pourrions employer aussi distinctement que nous connaissons les divers métiers de nos artisans, nous les pourrions employer en même façon à tous les usages auxquels ils sont propres, et ainsi nous rendre *maîtres et possesseurs de la nature* ». (Ce texte si célèbre se trouve dans le *Discours de la méthode,* sixième partie, intitulée : *Choses requises pour aller plus avant en la recherche de la nature.*)

C'est donc de manière artificielle que nous n'évoquerons pas ici le développement de la physique, renvoyant le lecteur à l'histoire de sa mathématisation. Dans le domaine strictement mathématique, nous nous limiterons, par ailleurs, à l'essentiel, c'est-à-dire à ce qui a fait époque, soit en renouvelant complètement une tradition ancienne soit en en créant une tout à fait nouvelle.

L'algèbre : Viète et la « logistique spécieuse »

Il faut alors, en premier lieu, rappeler le rôle essentiel qu'a joué le perfectionnement de l'outil algébrique. Une lente et progressive habituation au calcul, à l'algorithme qui développe le caractère opératoire de l'algèbre, conduit à isoler la notion d'opération.

Avec Viète (1540-1603) devient usuel le calcul sur des *symboles,* ainsi que l'idée de transformation algébrique qui conduit à celle d'une algèbre toute en formules, d'une méthode opératoire sur les *espèces* et *formes* des choses ou « logistique spécieuse ». C'est certainement en cela que consiste le progrès décisif de l'algèbre au XVIIe siècle, devenir une méthode sous les lois de laquelle se rangent tous les procédés médiévaux et italiens. Viète mit en évidence l'isomorphisme fondamental entre le domaine de l'algèbre numérique de Diophante, dont on venait de découvrir les travaux, de Cardan, de Tartaglia, de Bombelli, de Stifel, et celui de l'analyse géométrique sous-jacente aux exposés synthétiques d'Euclide, d'Archimède et d'Apollonios dont les œuvres récemment retrouvées de Pappus donnent une idée plus précise. Il subdivise l'analyse en trois parties : la première, la zététique ou art de chercher, consiste à adopter un symbolisme permettant de noter tant les grandeurs inconnues que les grandeurs connues, à exprimer les liens qui les unissent, à en dégager l'équation qui résume, de façon abstraite, le problème posé. La deuxième partie étudie, transforme et discute cette équation, et la troisième résout l'équation, soit par des constructions s'il s'agit de géométrie, soit par des calculs numériques s'il s'agit d'arithmétique. On y reconnaîtra les principes mêmes des méthodes modernes.

La théorie des équations est la première discipline à profiter des notations introduites par Viète et améliorées par Harriot (1560-1621). C'est Albert Girard (1595-1633) qui établit cette théorie sur les relations, mises en évidence par les méthodes de Viète, entre les coefficients et les racines, et dégage le principe qu'une équation a un nombre de racines égal à son degré. Descartes, de son côté, publie en 1637, en annexe au *Discours de la méthode,* sa *Géométrie,* dont le livre III expose la théorie des équations algébriques. Descartes fait la synthèse des résultats obtenus par ses contemporains et le bilan des progrès de l'algèbre en exposant, de façon théorique, la manière moderne de trouver les racines rationnelles d'une équation polynomiale. Mais c'est Newton qui reprendra la méthode effective de Viète pour la résolution numérique approchée.

Naissance de la géométrie analytique

En second lieu, c'est un fait justement célébré que la naissance de la géométrie analytique entre les mains de Descartes et Fermat, vers la même époque, et indépendamment l'un de l'autre. Elle a consisté en l'application de la nouvelle logistique spécieuse à l'analyse des lieux géométriques, en particulier des coniques, telle qu'elle apparaît chez Apollonios et Pappus. Mais Descartes s'est constitué son propre langage et sa propre notation, indépendamment de la tradition directe de Viète. « Tous les problèmes de géométrie se peuvent facilement réduire à tels termes, qu'il n'est besoin, écrit Descartes, que de connaître la longueur de quelques lignes droites, pour les construire. » Toute l'arithmétique repose, précise-t-il, sur les opérations arithmétiques usuelles, qu'il suffit d'introduire en géométrie; on peut ajouter une ligne *a*



S.E.F.

◀ **Descartes *est considéré, notamment, comme le fondateur de la géométrie analytique (Portrait de F. Hals, Paris, Louvre).***

à une ligne b et écrire $a + b$, ôter les lignes a et b l'une de l'autre et écrire $a — b$, etc. Descartes algébrise ainsi les opérations géométriques, et inversement, ce qui est encore plus important, donne une méthode pour interpréter géométriquement des expressions algébriques, si bien qu'à tout nombre correspond un segment, à toute opération un calcul segmentaire. C'est pourquoi il s'agit d'une invention éclatante, qui, selon l'expression de l'historien H. Zeuthen, fait passer la géométrie du stade de l'artisanat à celui de la grande industrie. Du fait que les concepts géométriques s'expriment algébriquement et les expressions algébriques s'interprètent géométriquement, les deux disciplines acquièrent une vitalité nouvelle. Mais cette « quantification » de la géométrie se fonde sur la priorité logique accordée à l'algèbre. Pourquoi ? Parce que, nous dit le cartésien Érasme Bartholin, si « dans les commencements il a été utile et nécessaire de donner des auxiliaires à notre faculté de spéculation pure : ce pourquoi les géomètres ont eu recours aux figures... », il est possible désormais de raisonner directement sur des lettres représentant des quantités abstraites. L'algèbre est donc une méthode ; c'est une *méthode de combinaison* qui permet, à partir des objets les plus simples, de progresser peu à peu dans la connaissance des objets les plus complexes. Tandis que la découverte par Fermat des coordonnées cartésiennes s'est limitée à une *répétition algébrique* de la mathématique grecque, Descartes fait de sa découverte une méthode d'invention d'une puissance et d'une universalité jusqu'alors inconnues en mathématiques : la mathématique grecque est définitivement dépassée. Il s'agit donc de la première rupture épistémologique dans l'histoire des mathématiques depuis l'invention hellène de la démonstration. Et au fond de cette rupture, il y a le *concept de fonction*.

L'apparition du concept de fonction

Si Descartes a énoncé clairement le principe de base de la géométrie analytique lorsqu'il résout le problème de Pappus, s'il a trouvé la définition précise des courbes géométriques, et montré que chaque point de celles-ci peut se construire, quelle que soit son abscisse, par une suite d'équations de degré de plus en plus grand, il a aussi banni de sa *Géométrie* toutes les courbes non susceptibles d'une définition analytique précise, en restreignant, en outre, les moyens de cette définition aux seules opérations algébriques. C'est pourquoi la géométrie analytique exclut de son domaine les courbes que Descartes appelle « mécaniques ». Si bien que, comme le souligne Bourbaki, Descartes reste en deçà de l'idée claire et féconde que Barrow (1630-1677) avait déjà du concept de fonction. A propos de la courbe définie par $x =$ cte, $y = f(t)$, avec l'hypothèse que $\frac{dy}{dt}$ soit croissante, Barrow dit expressément qu' « il n'importe en rien » que $\frac{dy}{dt}$ croisse « régulièrement suivant une loi quelconque, ou bien irrégulièrement », c'est-à-dire soit susceptible ou non d'une définition analytique. Il a fallu attendre le XIXe siècle pour que cette idée de Barrow soit précisée.

Car le disciple de Barrow, le grand Newton, a adopté le point de vue « algébrique » de Descartes. Leibniz lui-même, qui l'élargit, se contente de l'adjonction explicite des courbes exclues par Descartes et de la considération implicite des opérations analytiques alors usuelles. C'est le « saut » leibnizien dans l'analyse qui explique partiellement la polémique contre Descartes et son école. « J'ai montré, écrit Leibniz, combien la géométrie de Descartes est bornée, combien les problèmes les plus importants ne dépendent point des équations auxquelles cette géométrie se réduit, qu'il est ridicule de croire, comme Malebranche, que l'algèbre est la plus grande et la plus sublime des sciences. » En fait, Leibniz consacre une grande partie de ses efforts à méditer sur les séries. Autrement dit, il a une idée de la notion de convergence, et ce grâce à J. Gregory (1638-1675), qui fut le premier à en avoir conçu la nécessité.

Le calcul infinitésimal

C'est pourquoi, à l'exception des incursions de Fermat (1601-1665) dans la théorie des nombres (établissement de la méthode de descente infinie, du « petit » théorème de Fermat : $a^p = a$ (modulo p) pour p premier, et du « grand » théorème de Fermat : l'équation $x^n + y^n = z^n$ n'a pas de solutions rationnelles pour n entier supérieur à 2), à l'exception aussi des incursions de Desargues dans la géométrie projective (qui nous donnent le concept de point à l'infini sur une droite et le « théorème de Desargues »), les progrès essentiels en cette période concernent des notions étroitement liées à l'analyse. Les figures les plus marquantes restent, bien sûr, Newton et Leibniz. Tous deux héritent d'une tradition déjà suffisamment établie : celle de Galilée, de Cavalieri, puis Barrow, Wallis, puis Fermat, Roberval, etc., qui leur permet une synthèse où le calcul infinitésimal acquiert presque sa forme classique.

Newton (1643-1727) est d'abord l'élève de Barrow, auquel il succède comme professeur à Cambridge, en 1696. Sa formation doit par ailleurs à Descartes, dont il étudie la géométrie, à Wallis, à Mercator (1620-1687), familiers des calculs liés à des séries de puissances. Son apport à la mathématique ne se limite pas à son œuvre mathématique proprement dite. Sa physique ou *Philosophiae naturalis principia mathematica* résume tout le progrès géométrique des siècles passés. L'*Arithmetica universalis* (1707) et le *Tractatus de quadratura curvarum* développent amplement les techniques de la géométrie analytique ; le *Methodus fluxionum et serierum infinitarum,* paru en 1736, après sa mort, systématise des idées dont les premières remontent à 1665.

Le calcul infinitésimal de Newton est lié à l'étude mécanique du mouvement d'un point, et, plus précisément, à la détermination de la vitesse à un temps t donné, quand la longueur du segment parcouru est donnée, ou de la distance parcourue, à un temps t, quand c'est la vitesse qui est donnée. Le temps apparaît donc comme une variable indépendante : il s'agit non du temps physique, mais d'une *grandeur* dont la croissance uniforme représente et mesure le temps. Newton se sert des notions et des notations algébriques de Barrow : la distance parcourue est la « fluente » de la vitesse : $s = v$; et, inversement, la vitesse v est la « fluxion » de la distance s : $v = s$ (la « fluxion » est ce que Leibniz appelle le « quotient différentiel » d'une fonction ; la « fluente » l' « intégrale » d'une fonction). Bientôt (1676), Newton s'aperçoit qu'on peut réitérer le procédé qui lie la fluente à la fluxion, de façon à obtenir une suite de grandeurs dont chacune est la fluxion de celle qui la précède et la fluente de celle qui la suit ; la méthode algébrique s'applique ainsi à des relations fonctionnelles, d'une généralité plus grande que les grandeurs auxquelles la limitait Descartes.

▶ ***Figure 13 :*** *représentation schématique du problème du* **triangle caractéristique** *pour lequel* **Leibniz** *inventa son calcul des différences.*

Illustrons par un exemple la méthode des fluxions. Soit à calculer la fluxion de l'équation : $z^3 - zy^2 + a^2x - b^3 = 0$, où x, y et z dépendent de t et où a et b sont des constantes. On calcule successivement les fluxions de z, de y et de x :

(1) Pour $z^3 - zy^2$ on a : $3z^2\dot{z} - \dot{z}y^2$

(2) Pour $-zy^2$ on a : $-2zy\dot{y}$

(3) Pour a^2x on a : $a^2\dot{x}$

(4) Pour $-b^3$ on a : 0 d'où l'équation :

(5) $3z^2\dot{z} - \dot{z}y^2 - 2zy\dot{y} + a^2\dot{x} = 0$

En désignant par $\underline{0}$ une quantité très petite, différente de zéro, $\underline{0}\dot{z}$, $\underline{0}\dot{y}$ et $\underline{0}\dot{x}$ seront les *moments* des fluentes z, y et x. Les fluentes s'accroissent de leurs moments quand $\underline{t}$ s'accroît de $\underline{0}$; d'où :

$$(z + \underline{0}\dot{z})^3 - (z + \underline{0}\dot{z})\,(y + \underline{0}\dot{y})^2 + a^2\,(x + \underline{0}\dot{x}) - b^3 = 0.$$

De cette dernière équation, on soustrait l'équation donnée, puis on divise par $\underline{0}$ qui est différent de 0 ; on obtient :

$$3z^2\dot{z} + 3\,\underline{0}z\dot{z}^2 + \underline{0}^2\dot{z}^3 - \dot{z}y^2 - 2zy\dot{y} - 2\,\underline{0}\dot{z}y\dot{y} - \underline{0}z\dot{y}^2 - \underline{0}^2\dot{z}\dot{y}^2 + a^2\dot{x} = 0.$$

En faisant tendre $\underline{0}$ vers 0, il vient :

$$3z^2\dot{z} - \dot{z}y^2 - 2zy\dot{y} + a^2\dot{x} = 0.$$

On y reconnaîtra la définition de la dérivée d'une fonction :

$$\lim_{\substack{h \to 0 \\ h \neq 0}} \frac{f(z+h) - f(z)}{h}.$$

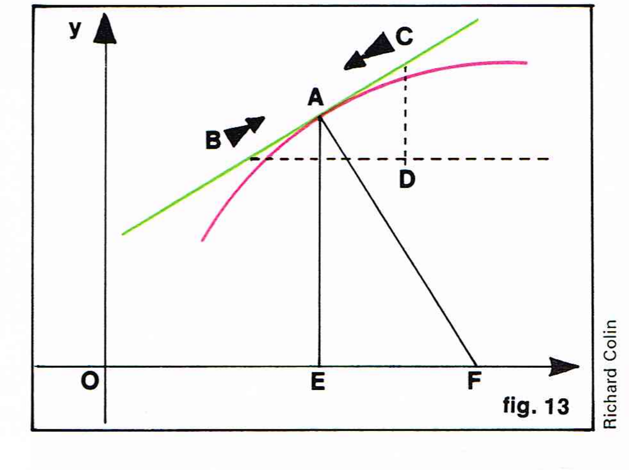
Ainsi Newton évite dans sa technique de démonstration aussi bien les anciennes démonstrations par l'absurde que la méthode de Cavalieri qui repose sur les indivisibles ; il introduit les « quantités évanouissantes », de manière à pouvoir assigner les *limites* assignables à des sommes et à des rapports de quantités déterminées prises à leur état de naissance ou d'évanouissement. Les quantités mises en relation dans l'expression de ces limites sont des éléments infinitésimaux de quantités finies, des *moments,* qui n'interviennent dans le calcul que par leur première proportion à la naissance.



▲ ***Le philosophe et mathématicien allemand*** **G. Leibniz** ***(1646-1716) ; il publia, en 1684, une*** **Nouvelle Méthode pour la détermination des maxima et des minima,** ***qui contient les principales lignes du calcul infinitésimal.***

De son côté, Leibniz, homme de système comme Descartes, conçoit, à la même époque que Newton, son *Calculus differentialis* et son *Calculus summatorius* pour résoudre deux problèmes distincts et, pour lui, inverses l'un de l'autre : celui des tangentes et celui des sommations (c'est-à-dire de l'intégration). Ses conceptions sont exposées dans deux mémoires publiés dans les *Acta eruditorum* (1684 et 1686) et dans la correspondance qu'il entretint avec les savants de son époque.

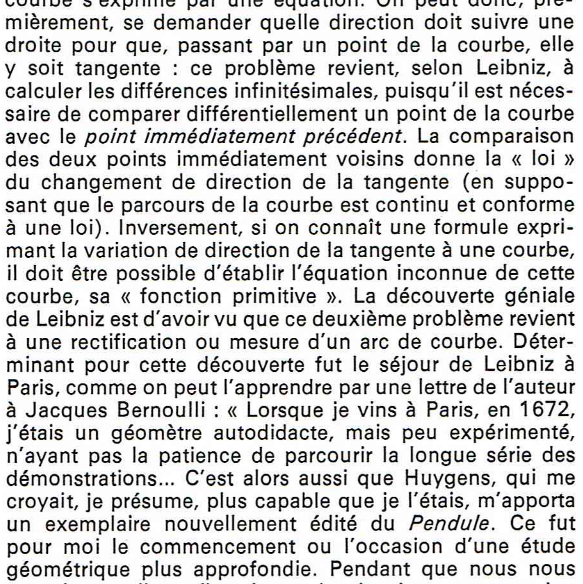
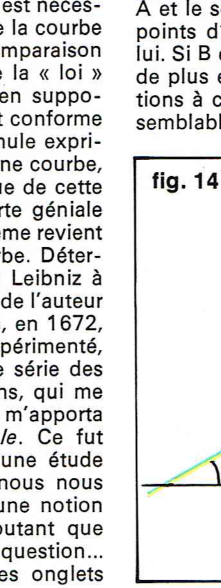
Dans un système de coordonnées cartésiennes, toute courbe s'exprime par une équation. On peut donc, premièrement, se demander quelle direction doit suivre une droite pour que, passant par un point de la courbe, elle y soit tangente : ce problème revient, selon Leibniz, à calculer les différences infinitésimales, puisqu'il est nécessaire de comparer différentiellement un point de la courbe avec le *point immédiatement précédent.* La comparaison des deux points immédiatement voisins donne la « loi » du changement de direction de la tangente (en supposant que le parcours de la courbe est continu et conforme à une loi). Inversement, si on connaît une formule exprimant la variation de direction de la tangente à une courbe, il doit être possible d'établir l'équation inconnue de cette courbe, sa « fonction primitive ». La découverte géniale de Leibniz est d'avoir vu que ce deuxième problème revient à une rectification ou mesure d'un arc de courbe. Déterminant pour cette découverte fut le séjour de Leibniz à Paris, comme on peut l'apprendre par une lettre de l'auteur à Jacques Bernoulli : « Lorsque je vins à Paris, en 1672, j'étais un géomètre autodidacte, mais peu expérimenté, n'ayant pas la patience de parcourir la longue série des démonstrations... C'est alors aussi que Huygens, qui me croyait, je présume, plus capable que je l'étais, m'apporta un exemplaire nouvellement édité du *Pendule.* Ce fut pour moi le commencement ou l'occasion d'une étude géométrique plus approfondie. Pendant que nous nous entretînmes, il me fit voir que je n'avais pas une notion assez exacte des centres de gravité... en ajoutant que Dettonville (Pascal) avait très bien traité cette question... Sans aucun retard, j'étudiais ces produits, ces onglets



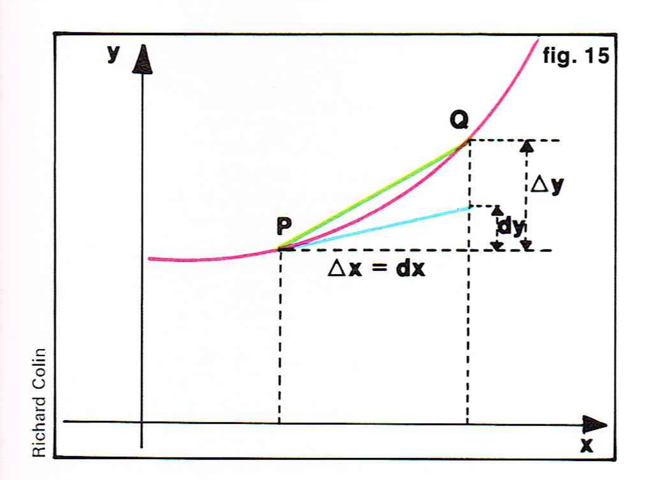
inventés par Vincent et perfectionnés par Pascal. Je voyais avec plaisir ces sommes, et les sommes de ces sommes, les solides qui en naissaient et leurs démonstrations. Tout cela me donnait plus de plaisir que de travail. J'en étais là lorsque, par hasard, je tombai sur une démonstration de Dettonville, très facile de son espèce... Mais quel fut mon étonnement de voir que Pascal avait eu les yeux fermés comme par un sort : car je vis aussitôt que le théorème pouvait s'appliquer généralement à toutes les courbes, bien que les perpendiculaires ne se rencontrassent pas dans un même centre... Je m'en vais aussitôt chez Huygens... et je lui exposai mon théorème général sur la rectification des courbes. Il fut saisi d'étonnement et me dit que c'était là le théorème sur lequel s'appuyaient ses constructions pour trouver les aires des conoïdes parabolique, elliptique et hyperbolique. » (Cité dans le tome II de *l'Histoire générale des sciences et des techniques,* P. U. F., pp. 235-236.)

Ainsi l'impulsion de Huygens et ses conseils, la lecture de Pascal *(Traité des sinus du quart de cercle),* l'étude des géomètres contemporains ont conduit Leibniz à voir dans l'étude du *triangle caractéristique* l'outil essentiel pour résoudre le problème des tangentes pour des courbes quelconques, puis à concevoir la rectification des courbes comme un problème inverse du premier. Il introduit la notation encore usitée de nos jours, soit, pour l'intégrale : $\int y\,dy$, et pour la différentielle : dx, $\int$ désignant une somme, tandis que d représente une différence.

Voyons sur un exemple comment raisonnait Leibniz (*fig.* 13). Considérons une droite tangente à une courbe en un point A, qui s'écarte de cette courbe aux points B et C ; B et C peuvent être les extrémités de l'hypothénuse du triangle rectangle dont les côtés de l'angle droit sont parallèles aux axes de coordonnées. Ce triangle est semblable au triangle formé par la normale à la courbe au point A, la parallèle à l'axe des ordonnées passant par A et le segment déterminé sur l'axe des abscisses par les points d'intersection de ces deux dernières droites avec lui. Si B et C se meuvent vers A, le triangle BDC deviendra de plus en plus petit, tout en gardant les mêmes proportions à cause de la continuité des lignes, et restera donc semblable au triangle AEF ; celui-ci, reproduction très

▶ ***Figure 14 :*** *voir démonstration dans le texte page* **299.**



◀ *Figure 15 : représentation schématique illustrant la distinction entre une différence et une différentielle.*

grande du triangle BDC, est le triangle « caractéristique » ; la loi des tangentes pourra être exprimée par un rapport entre les côtés du triangle caractéristique, par une *fonction* de l'angle d'inclinaison de la tangente sur l'un des axes de coordonnées. D'où la nécessité d'établir l'équation de cette fonction et son rapport avec l'équation de la courbe, ce pour quoi Leibniz invente son calcul des différences et, par passage à la limite, le calcul différentiel.

Pour un accroissement Δx de la valeur de l'abscisse, l'ordonnée y s'accroît de Δy (*fig.* 14). Dans le cas d'un accroissement fini, on considère l'angle α que fait la sécante à la courbe passant par les points P et R, avec l'axe des abscisses, et on obtient la tangente de α comme quotient de Δy par Δx. Mais si $y = f(x)$,

$$\Delta y = f(x + \Delta x) - f(x),$$

d'où
$$\frac{\Delta y}{\Delta x} = \frac{f(x + \Delta x) - f(x)}{\Delta x}.$$

Ce rapport donne donc la tangente trigonométrique de l'angle α. Et comme on s'intéresse surtout à la tangente au point P, l'accroissement de la variable x doit avoir une valeur infiniment petite dx, à laquelle correspond un accroissement dy de y ; on procède comme dans le cas des différences finies, en écrivant le rapport :

$$\frac{dy}{dx} = \operatorname{tg} \alpha = \frac{f(x + dx) - f(x)}{dx}.$$

Le passage du cas fini au cas des accroissements infiniment petits s'exprimerait aujourd'hui par l'équation :

$$\frac{dy}{dx} = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}$$

Leibniz ne distinguait pas par l'écriture Δy de dy, bien qu'il fît parfaitement la distinction entre une différence et une différentielle, distinction que le lecteur saisira aisément en se reportant à la *figure* 15.

Telle est l'allure de l'invention leibnizienne qui couronne, avec celle de Newton, les découvertes successives du XVII[e] siècle. L'aboutissement en est la création de l'analyse du XVIII[e] siècle dans sa double relation aux problèmes de physique mathématique d'une part, et d'autre part, à travers les développements donnés par les frères Bernoulli au calcul leibnizien, à la résolution des équations différentielles.

Le XVIII[e] siècle

La richesse de la grande invention du XVII[e] siècle apparaît au siècle suivant dans ses applications à la physique et à la géométrie. Le siècle des lumières foisonne en découvertes ; les premières sont dues à de Moivre, Taylor et Stirling, pour ne citer que les noms les plus illustres de l'école britannique, et aux frères Bernoulli sur le continent ; puis le siècle culmine avec l'œuvre de Leonhard Euler avant de s'achever par la publication des grands traités de l'École de Paris par Lagrange, Laplace, Monge, Legendre et Lacroix.

Les Bernoulli

Jacques Bernoulli (1654-1705) s'initia au calcul de Leibniz dont il défendit et propagea les méthodes à la fois dans son œuvre et dans son enseignement à l'université de Bâle. Il est resté célèbre par l'introduction, à l'occasion de la sommation des puissances p-ièmes des premiers entiers, des nombres dits « de Bernoulli », par une théorie des probabilités contenue dans son *Ars conjectandi,* par le recours au raisonnement par récurrence, etc. Dans ses travaux sur les séries, il établit la divergence de la série harmonique ; il résolut le problème de la ligne isochrone, posa celui de la chaînette, résolu par Huygens, Leibniz et Jean Bernoulli. Il étudia les courbes élastiques et s'attaqua à de nombreuses autres questions, dont certaines ont conduit à une problématique de géométrie infinitésimale, d'autres, comme celle des isopérimètres, à l'ébauche du calcul des variations.

Jean Bernoulli (1667-1748), nommé d'abord professeur à Groningen, succéda à son frère Daniel à Bâle, en 1705. On lui doit les premiers manuels de *Calcul différentiel et intégral,* qui consignent ses cours de l'année 1691-1692. Par ailleurs, il étudia l'exponentielle $y = x^x = e^{x \operatorname{Log} x}$, $(x > 0)$, la série « de Bernoulli » :

$$\int_0^x y dx = xy - \frac{1}{2!} x^2 y' + \frac{1}{3!} x^3 y'' + ...,$$

la chaînette, les lignes isochrone et brachistochrone, et la cycloïde ; il fit le premier emploi des coordonnées polaires et une étude systématique des équations différentielles du premier ordre, en recourant notamment aux méthodes de séparation des variables et de variation des constantes. Enfin, il contribua à la formation de Leonhard Euler. On sait, d'autre part, que le marquis de L'Hôpital doit à Jean Bernoulli le contenu de son livre sur l'*Analyse des infiniment petits*.

Outre Daniel Bernoulli (1700-1782), le fils de Jean, il faut aussi mentionner A. Claude Clairaut, qui appliqua l'analyse à des problèmes géométriques et prit part à l'étude des équations différentielles.

L'école anglaise

Les noms de Taylor (1685-1731) et de Maclaurin (1698-1746) sont indissolublement liés à la théorie des séries. La célèbre formule de Taylor se trouve dans sa *Methodus incrementorum directa et inversa* (1715) :

$$f(x + h) = f(x) + h \cdot f'(x) + \frac{h^2}{2!} f''(x) + ... ;$$

tandis que celle de Maclaurin en explicite un cas particulier, celui où $x = 0$ et $h = x$.

Stirling (1692-1770) est surtout connu pour sa formule d'approximation de $n!$ (pour n entier très grand) :

$$n! \cong n^n \cdot e^{-n} \sqrt{2\pi n}.$$

Abraham de Moivre (1667-1754) se signale par la série :

$$\operatorname{Log} n! = \operatorname{Log} \sqrt{2\pi n} + n \cdot \operatorname{Log} n - n + \sum_{k=1}^{\infty} \frac{B_{2k} \cdot n^{1-2k}}{2k(2k-1)},$$



◀ *Le mathématicien* **Abraham de Moivre** *(1667-1754) introduisit la trigonométrie des quantités imaginaires et donna dans ce domaine, en 1730, la formule qui porte son nom.*

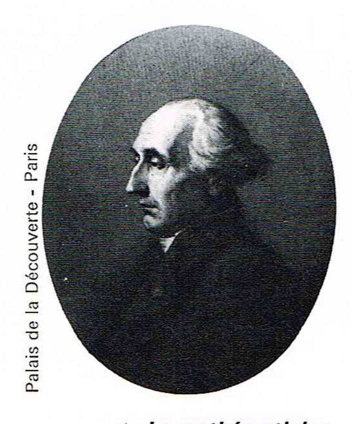
où les coefficients B_{2k} sont les nombres de Bernoulli, tandis que la formule à laquelle il a laissé son nom :

$$(\cos x + i \sin x)^n = \cos nx + i \sin nx$$

est en réalité due (sous sa forme connue) à Euler.

Les travaux d'Euler

Leonhard Euler (1707-1783) est un des mathématiciens les plus féconds et les plus profonds de tous les temps. Fils d'un pasteur de Bâle, il étudie d'abord la théologie et se découvre une vocation de mathématicien en écoutant les cours de Jean Bernoulli. On le trouve en 1727 à l'Académie de Saint-Pétersbourg, où il est professeur de physique en 1730 et de mathématiques en 1733. Appelé en 1741 à l'Académie de Berlin, il y dirige en 1745 le département de mathématiques; il rencontre là Maupertuis et d'Alembert. Euler est important à la fois par son génie créateur et par ses manuels, qui exercèrent une grande influence; l'*Introductio in analysin infinitorum* (1748) est le plus célèbre de ses ouvrages, et le premier traité classique d'analyse. Euler y fait une étude systématique des fonctions élémentaires, qu'il classe suivant leur mode de formation. Une fonction est définie comme une expression analytique formée d'une manière quelconque à partir d'une quantité variable et de constantes; les expressions polynomiales, les séries entières, les expressions logarithmiques ou trigonométriques sont permises, ce qui autorise la considération de fonctions de plusieurs variables. Il distingue (comme Jean Bernoulli) les fonctions algébriques et les fonctions transcendantes, ces dernières pouvant être engendrées par des séries infinies, puis les fonctions explicites et les fonctions implicites, enfin les fonctions uniformes et les fonctions multiformes. En 1749, à propos de l'équation aux dérivées partielles des cordes vibrantes, Euler est conduit à élargir sa définition du concept de fonction de manière à ne plus exclure les courbes tracées au hasard sur le plan. Sa classification a le mérite d'avoir mis au centre de l'organisation de l'édifice analytique ce concept de fonction, en facilitant ainsi le travail critique qui s'appliquera ultérieurement à redéfinir les notions fondamentales de l'analyse.

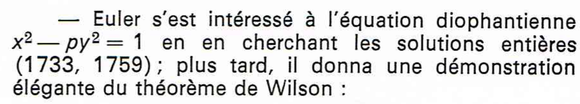


Palais de la Découverte - Paris

▲ *Le mathématicien* **L. de Lagrange** ***(1736-1813).***

Les *Institutiones calculi differentialis* (1755) montrent un égal souci de rigueur. Contrairement à Maclaurin, par exemple, qui tenta d'asseoir le calcul infinitésimal sur la géométrie, et en particulier sur la méthode d'exhaustion telle qu'elle est utilisée par Archimède, Euler se tourne plutôt vers la manipulation formelle des expressions algébriques, ce qui le conduit à identifier les quantités infiniment petites ou évanouissantes à de simples zéros; les différentielles *dx* et *dy* sont considérées comme nulles, puisqu'elles n'ont pas de *grandeur assignable*, et le problème se pose d'expliquer que le rapport $\frac{dy}{dx}$, qui a la forme indéterminée $\frac{0}{0}$, puisse égaler une quantité déterminée. Euler pense que, puisque $n \cdot 0 = 0$ pour tout nombre *n*, on peut avoir $n = \frac{0}{0}$, et le calcul de la dérivée de la fonction considérée sert à déterminer précisément ce rapport. $(dx)^2$ s'évanouit *avant dx*, en sorte que le rapport de $dx + (dx)^2$ à *dx* est égal à 1. Euler est donc conduit à distinguer différents *ordres d'infini*, bien qu'on ne disposât alors que de l'unique symbole ∞ : $\frac{a}{0}$ est un infini de premier ordre, tandis que $\frac{a}{(dx)^2}$ est un infini de second ordre, etc.

Les *Institutiones calculi integralis* (1768) constituent, avec l'ouvrage précédent, une somme de résultats et une référence irremplaçable avant la publication par Lacroix des deux volumes de son *Traité de calcul différentiel et intégral* (Paris, 1797-1800). Euler y étudie l'intégration des fonctions rationnelles, irrationnelles et transcendantes élémentaires (l'intégration étant considérée comme l'opération inverse de la dérivation); les équations différentielles sont intégrées par la technique de séparation des variables ou par le biais des développements en série; des indications sont données sur l'intégrale elliptique et le calcul des variations. Souvent indécis sur les concepts fondamentaux de l'analyse, les manuels d'Euler fourmillent de trouvailles, dont il nous faut citer les plus importantes pour faire le point.



— Euler s'est intéressé à l'équation diophantienne $x^2 - py^2 = 1$ en en cherchant les solutions entières (1733, 1759); plus tard, il donna une démonstration élégante du théorème de Wilson :

$$(p - 1)! = -1 \pmod p.$$

— Il a donné au petit théorème de Fermat une expression plus générale grâce à l'emploi de la fonction dite d'Euler : $\varphi(n)$ étant le nombre des entiers inférieurs à *n* et premiers avec lui.

— Il a montré l'irrationalité des nombres *e* et e^2; c'est, du reste, à lui que nous devons cette notation *e*, ainsi que la vulgarisation des symboles *i* et π.

— Il a établi la divergence de la série $\sum_{n=1}^{\infty} \frac{1}{p^n}$ (*p* premier) et donné une valeur pour la fameuse *constante d'Euler* :

$$C = \lim_{n \to \infty} \left(1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{n} - \text{Log } n\right) = 0{,}577\,2$$

— Il a établi la formule : $e^{ix} = \cos x + i \sin x$.

— Les intégrales d'Euler :

$$\int_0^1 x^p (1 - x)^q \, dx \quad \text{et} \quad \int_0^1 x^p \cdot e^{-x} \, dx$$

(*p* et *q* entiers positifs) sont bien connues aujourd'hui sous les noms de fonction β et fonction Γ.

— Euler, enfin, ouvre un nouveau chapitre dans l'arithmétique, en étendant la notion de divisibilité aux entiers d'une extension quadratique quelconque.

L'école française

L'œuvre immense d'Euler, parachèvement de celle de Leibniz, fut décisive dans la préparation du courant critique qui se développe déjà dans la réflexion de D'Alembert et les doutes qu'il exprime sur l'usage des séries non convergentes. D'Alembert, en effet, s'est efforcé, dans les articles *Différentiel* et *Limite* de l'*Encyclopédie* et dans les *Éclaircissements sur les éléments de philosophie* (1767), d'exposer les propriétés des infiniment petits d'une manière systématique et claire et de montrer que toute la « métaphysique » du calcul infinitésimal se réduit aux concepts de dérivée et de limite.

L. de Lagrange (Turin, 1736 - Paris, 1813) exerça une influence décisive sur la tradition analytique française qu'il contribua à créer par son enseignement, d'abord à l'École normale de l'an III, puis à l'École polytechnique (Lagrange, d'abord professeur à l'école d'artillerie de Turin, avait succédé à Euler à l'Académie de Berlin; c'est en 1787 qu'il se fixe à Paris). De cet enseignement il a tiré la matière de ses ouvrages essentiels : la *Mécanique analytique* (1788), la *Théorie des fonctions analytiques* (1797) et les *Leçons sur le calcul des fonctions* (1799). Ayant traité de domaines variés qu'il a souvent renouvelés, tels que les équations différentielles, le calcul des variations, etc., il s'est également illustré par ses mémoires algébriques qui inaugurent une période décisive pour la théorie des équations algébriques. Dans le mémoire de 1771 *(Réflexions sur la résolution algébrique des équations),* il écrit : « Je me propose d'examiner les différentes méthodes que l'on a trouvées jusqu'à présent pour la résolution algébrique des équations, de les réduire à des principes généraux et de faire voir *a priori* pourquoi ces méthodes réussissent pour le troisième et le quatrième degré, et sont en défaut pour les degrés ultérieurs. » Lagrange s'intéresse ainsi aux méthodes plus qu'aux équations elles-mêmes, d'une manière *a priori,* expliquant réussites et échecs des tentatives antérieures pour établir des formules de résolution. Bref, comme l'a écrit Bourbaki, Lagrange procède à « une analyse systématique des problèmes posés et des méthodes susceptibles de les résoudre, analyse qui, en soixante ans, conduira aux résultats définitifs de Galois ».

Pierre-Simon de Laplace (1749-1827) applique les procédés de l'analyse aux probabilités, ce qui le conduit à étudier l'intégrale $\int_0^{\infty} e^{-u^2} \, du$. La *Théorie analytique des probabilités* (1812) donne une synthèse de toute l'œuvre du siècle en ce domaine. Laplace s'est également intéressé aux équations aux dérivées partielles, en donnant

Giraudon

son nom à l'équation $\sum_{i=1}^{2} \frac{d^2\varphi}{dx_i^2} = \Delta\varphi = 0$, qui jouera un rôle important dans la théorie des fonctions, et surtout à la *Mécanique céleste,* à laquelle il a consacré un traité monumental en cinq volumes (1799-1825) qui couronne les travaux de Newton, Clairaut, d'Alembert, Euler et Lagrange sur le système des planètes, la forme de la Terre, la gravitation et le problème des trois corps.

Le dernier des trois grands analystes français, Adrien-Marie Le Gendre (1752-1833), appartient déjà au début du XIXe siècle par ses traités sur la géométrie (1794), sur la théorie des nombres (1798), sur les fonctions elliptiques et sur les intégrales eulériennes (1825-1832), ainsi que par ses exercices de calcul intégral. Ses premiers travaux, qui datent d'avant la coupure de 1789, concernaient le calcul des variations, les équations aux dérivées partielles et la géodésie.

En cette fin du XVIIIe siècle, outre la percée algébrique de Lagrange dont nous avons parlé plus haut, le phénomène le plus important est sans doute l'application par Gaspard Monge (1746-1818) de l'analyse à la géométrie et à la mécanique. Monge crée la *Géométrie descriptive* (1794-1795), ouvre la voie, avec son élève Poncelet, à la géométrie projective complexe. Aussi Bourbaki reconnaît-il « sous le nom de "principe des relations contingentes" chez Monge, ou de "principe de continuité" chez Poncelet, le premier germe de l'idée de "spécialisation" de la géométrie algébrique moderne ». Notons aussi que les *Feuilles d'analyse appliquée à la géométrie* (1795) de Monge préparent à la naissance, au siècle suivant, de la géométrie différentielle.

Nous n'avons pu évoquer ici tous les travaux mathématiques, non plus que tous les mathématiciens du XVIIIe siècle : nous n'avons rien dit, par exemple, d'A. de Vandermonde, de Lambert (1728-1777), de Waring (1734-1793), des autres membres de la dynastie des Bernoulli. C'est que, faute de pouvoir retenir tous les détails, nous avons voulu surtout souligner la percée des mathématiques vers des formes inédites : vers une réflexion sur ses concepts fondamentaux que le XIXe siècle développera pour mettre de l'ordre dans le foisonnement des découvertes accumulées.



◀ *Le mathématicien français* **Gaspard Monge** *(1746-1818), dont les ouvrages et l'enseignement oral eurent une influence considérable sur tous les mathématiciens du XIXe siècle.*

Giraudon

Le XIXe siècle

Le XIXe siècle mathématique nous montre une discipline en plein essor : multiplication des travaux, extension et diversification des recherches. Trois traits cependant dominent cette période :

— le renouvellement profond de disciplines bien connues, consécutif au souci de rigueur logique et à l'esprit d'abstraction qui avaient commencé à poindre au siècle précédent : ainsi en est-il de l'algèbre avec Gauss et Galois, de l'analyse avec Cauchy, Bolzano et Weierstrass, et de l'arithmétique, à la fin du siècle ;

— la conquête de domaines absolument nouveaux par des disciplines inédites : théorie des groupes, théorie des fonctions ou théorie des ensembles (nous ne dirons rien de l'histoire de cette discipline, devenue le centre des mathématiques contemporaines ; on se reportera au chapitre qui lui est consacré) ;

— le rapide développement de la physique mathématique qui profite de l'outil mathématique, mais pose, en retour, des problèmes féconds dont l'étude oriente l'évolution de certaines branches.

Des facteurs politiques et socio-économiques commandent cette éclosion. Par sa réforme de l'enseignement supérieur scientifique et technique, la Révolution française accorde aux mathématiques une place bien plus grande que par le passé dans les programmes scolaires et dans la société (voir *Mathématiques et société*). On sait le rôle joué par l'École polytechnique d'abord, l'École normale supérieure ensuite, dans la formation de mathématiciens de premier ordre. Des revues spécialisées voient le jour : *Journal de l'École polytechnique* (1795), *Annales de mathématiques* de J.-D. Gergonne (1811-1832), *Journal de mathématiques pures et appliquées* de J. Liouville (1837), etc. ; et les *Comptes rendus de l'Académie des sciences,* fondés en 1835, assurent la diffusion rapide des résultats nouveaux. A la faveur des nationalismes politiques et des progrès de l'industrie, ce mouvement s'étend à d'autres pays d'Europe ou d'Occident : en Allemagne, où les universités de Göttingen, de Berlin, de Königsberg, Bonn, Halle, etc., disputent rapidement à Paris le privilège de centre de recherche et de rayonnement, et où est édité, à partir de 1826, le célèbre *Journal de Crelle (Journal für die reine und angewandte Mathematik) ;* en Angleterre, en Italie, en Russie même. Un peu partout sont fondées des sociétés mathématiques : *London Mathematical Society* (1865), *Société mathématique de France* (1872), *Circolo matematico di Palermo* (1884), *American Mathematical Society* (1888), etc.

Le premier congrès international de mathématiques eut lieu à Zürich en 1897.

La richesse et la diversité des acquis sont telles que nous commençons l'histoire de ce siècle par un tableau (*tableau* V) qui ne prétend pas à l'exhaustivité, mais où le lecteur trouvera les repères indispensables.

Le renouveau de l'algèbre

Théorie des équations et théorie des groupes

En algèbre, comme en de nombreux autres secteurs, le siècle est dominé, à son début, par le génie et la puissance de Carl Friedrich Gauss (1777-1855). C'est lui qui donne, dès sa thèse (1799), la preuve du théorème fondamental de l'algèbre, selon lequel tout polynôme de degré n a exactement n racines. Ce théorème avait été énoncé pour la première fois en 1629 par Girard, puis démontré, mais de façon imparfaite (en 1746), par D'Alembert et Euler. La démonstration de Gauss, qui utilise des résultats d'analyse alors connus, présuppose l'existence des nombres complexes.

Mais au début de ce XIXe siècle, le problème principal demeure celui de la résolubilité des équations de degré supérieur à 4. Sa solution, amorcée dans les travaux de Lagrange et de Vandermonde, interviendra avec le développement de la théorie des groupes et de la théorie des corps. C'est dans cette perspective que s'inscrivent l'effort de Gauss relatif à la résolution de l'équation $x^n - 1 = 0$, qui lui fait pressentir la notion de groupe cyclique, et l'étude de Paolo Ruffini (1765-1822) du comportement des fonctions rationnelles des racines lors des permutations de celles-ci, par laquelle il tente de prouver l'impossibilité de la résolution de l'équation générale du 5e degré.

Tableau V	
Avant 1800	**Après 1800**
On connaît les différentes sortes d'opérations et de nombres.	**On a le concept de nombre réel, la théorie des nombres complexes et hypercomplexes, le calcul vectoriel.**
On a une théorie élémentaire des nombres.	**On dispose de formes quadratiques, de la théorie des nombres algébriques, et d'une théorie analytique des nombres.**
L'algèbre est surtout développée comme théorie des équations.	**L'algèbre gagne en généralité; les théories des groupes et des invariants font leur apparition.**
Éléments d'analyse avec des applications à la géométrie et à la physique.	**L'analyse classique se constitue et s'organise autour de concepts précisément définis. Théorie des fonctions et géométrie différentielle se développent comme des branches autonomes.**
Manipulation des séries dans le calcul différentiel et intégral, surtout des séries entières.	**Théorie rigoureuse de la convergence. Invention des séries de Fourier.**
Géométrie analytique.	**Algèbre linéaire.**
Géométrie descriptive.	**Géométrie projective, topologie, théorie des ensembles, logique mathématique, telles sont les dernières créations du siècle.**

▲ ***Tableau V : les grands moments mathématiques du XIXe siècle.***

Après Abel (1802-1829), qui montre plus rigoureusement que Ruffini l'impossibilité de résoudre par radicaux l'équation générale du 5e degré, et recherche les critères caractérisant les équations résolubles par radicaux, c'est Évariste Galois (1811-1832) qui présente à l'Académie des sciences un mémoire sur ce sujet (1831). A la base de la théorie de Galois, qui ne fut vraiment connue qu'après que Liouville eut publié, en 1846, les œuvres de ce génie tué à 21 ans dans un duel, on trouve les notions de corps, esquissées par Gauss en 1801, d'adjonction et de polynôme irréductible sur un corps donné, qui seront développées par Riemann et Dedekind. On y trouve aussi les principes de la théorie des groupes de substitution.

L'idée fondamentale de la théorie de Galois est qu'une équation algébrique est résoluble par radicaux si et seulement si un certain groupe, le groupe de Galois de l'équation, possède la propriété d'être « résoluble ». Or, le groupe de Galois d'une équation est isomorphe à un groupe de permutations de ses racines, si bien qu'on découvrira les propriétés d'une équation à coefficients sur un corps K en étudiant l'ensemble des permutations de certains éléments du plus petit sur-corps de K, L, obtenu à partir de K par adjonction des racines de l'équation.

La théorie de Galois permet, en particulier, de résoudre les problèmes, posés depuis l'Antiquité, de construction à la règle et au compas : ni la trisection de l'angle, ni la duplication du cube, ni la quadrature du cercle ne sont possibles à la règle et au compas.

A. Cayley appliqua la théorie des groupes aux quaternions (1854), et Hamilton étudia les groupes de symétrie des polyèdres réguliers; mais ce n'est qu'après le *Traité des substitutions* de Camille Jordan (1838-1922) que l'importance des idées de Galois apparut en plein jour et que la théorie des groupes intervint dans les secteurs les plus divers (de la géométrie aux équations différentielles). C'est par le biais de la notion de groupe qui révèle ainsi, sous la diversité des représentations et des langages, l'identité des lois et des opérations, que l'idée de structure abstraite fit sa première entrée dans les mathématiques modernes. Après 1870, Sophus Lie et Felix Klein profiteront de ce nouvel outil pour faire ressortir, dans la diversité des théories, l'unité structurelle de la géométrie.

L'algèbre linéaire - Les algèbres

Un des traits caractéristiques du XIXe siècle mathématique est le relief de plus en plus marqué que prennent les *problèmes linéaires :* étude des déterminants, introduction des matrices, étude des formes algébriques et des invariants, théorie des quaternions et des nombres hypercomplexes, introduction de nouveaux types d'algèbres. L'algèbre linéaire proprement dite, d'abord limitée à l'étude des systèmes d'équations algébriques du premier degré, s'étend aux systèmes d'équations différentielles et aux dérivées partielles, et sa force d'explication trouve une heureuse application en géométrie.

Nous devons à Cauchy l'introduction du terme « déterminant », dans son sens moderne, à Cayley notre notation actuelle, mais ce sont les travaux de Jacobi (1795-1855) qui posent les principes de la théorie générale des déterminants et contribuent à la diffusion de ce nouvel algorithme.

Extensions du concept de déterminant, les matrices apparaissent dans les études sur la composition homographique de Cayley, dès 1843, et sont précisément définies par lui en 1858. Entre-temps, Hamilton les a également introduites dans ses *Lectures on Quaternions* (1853), et le calcul géométrique de Grassmann les utilise implicitement. Le succès du calcul matriciel est remarquable à la fin du siècle, dans l'école anglaise, avec Clifford et Sylvester, puis en Amérique, avec Benjamin Peirce qui l'utilise dans sa théorie des algèbres linéaires associatives.

Le *concept d'invariant* est tout aussi important et tout aussi distinctif que celui de groupe; lui aussi permet l'unification de questions algébriques et géométriques. Il s'impose dans l'étude des formes algébriques, c'est-à-dire des fonctions homogènes à plusieurs variables. Dans un système de coordonnées homogènes, l'équation d'une courbe plane, ou d'une surface, se ramène à l'annulation d'une forme binaire, ou ternaire, et les changements de coordonnées reviennent à des substitutions. L'étude des propriétés de figures équivaut donc analytiquement à celle des propriétés des formes et conduit à leur réduction à leur forme canonique et à la recherche de leurs invariants et covariants, c'est-à-dire aux fonctions de leurs coefficients qui ne sont pas affectées par certaines transformations. Le concept d'invariant, sous-jacent à divers travaux de Lagrange, de Gauss, Cauchy, Jacobi, Eisenstein, est explicite, en 1841, chez Boole; le terme apparaît chez Sylvester (1814-1897) qui est, avec Cayley (1821-1895), l'auteur des fondements de la théorie des formes algébriques et des invariants. Signalons que Sylvester transmit le patrimoine mathématique européen aux États-Unis, où il fut professeur et fonda l'*American Journal of Mathematics*.

La théorie des formes algébriques et des invariants, qui avait reçu une impulsion de l'étude analytique des propriétés projectives des courbes et des surfaces algébriques, lui fournit en retour un langage commode et une méthode analytique de découverte. Alfred Clebsch (1833-1872), fondateur (avec Carl Gottfried Neumann) des *Mathematische Annalen* (1868), créa, avec Siegfried Aronhold (1819-1884), une symbolique nouvelle qui conduisit à l'algèbre tensorielle. Clebsch applique la théorie des invariants à des questions de géométrie projective. Son utilité fut grande, également, en théorie des nombres, des équations différentielles et aux dérivées partielles; Lie et Klein en firent ressortir les liens avec la théorie des groupes. Hilbert, enfin, dans l'un de ses premiers travaux (1890), réussit à en dégager les lois fondamentales d'une façon particulièrement concise et élégante, en étendant le théorème de finitude aux formes algébriques à n variables.

Quaternions et vecteurs

Après que Gauss eut justifié l'existence des nombres complexes en en donnant une interprétation géométrique, on tenta de définir des nombres complexes de type $ai + bj + ck$, mais on s'aperçut qu'on ne pouvait définir de façon univoque une multiplication pour de telles

sommes de trois facteurs. W. R. Hamilton (1805-1865) et H. Grassmann (1809-1877) inventèrent, le premier les « quaternions », le second les « grandeurs extensives » dont le rôle fut décisif dans le développement du calcul vectoriel et du calcul tensoriel.

Hamilton publia en 1835 sa *Theory of Algebraic Couples* qui contient une représentation algébrique des nombres complexes comme couples (ordonnés) de nombres réels :

$$z = a + bi = (a, b).$$

La multiplication, ainsi définie :
$(a, b) \cdot (c, d) = (ac - bd, ad + bc)$, redonne l'égalité $i^2 = -1$, quand on donne à a et à c la valeur 0, et à b et à d la valeur 1. Les « quaternions », que Hamilton découvre dès 1843, apparaissent dans les *Lectures on Quaternions* (1853) et dans les *Elements of Quaternions* (ouvrage posthume), sous la forme : $t + ix + jy + zk$, où t est la partie scalaire, et $ix + jy + zk$ la partie vectorielle. Les quaternions sont avec les nombres complexes les seuls systèmes de nombres à coefficients réels sur lesquels la multiplication est associative et la division possible.

En 1844, H. Grassmann publia *Die lineale Ausdehnungslehre, ein neuer Zweig der Mathematik,* où il introduisit ses « grandeurs extensives » d'une façon plus générale que n'avait fait Hamilton pour les quaternions; en effet, il se place d'emblée dans un espace à n dimensions en définissant un vecteur par la formule : $V = \sum_{i=1}^{n} x_i n_i$ et distingue le produit scalaire (intérieur) du produit vectoriel (extérieur), ce dernier n'étant pas commutatif. L'ouvrage de Grassmann, d'un abord difficile, ne suscita de l'intérêt que lorsque W. G. Hankel (1867) et Schlegel (1872-75) en eurent donné une présentation plus claire.

Quaternions de Hamilton et « grandeurs extensives » de Grassmann correspondent à des extensions du concept de nombre et entraînent un élargissement de la notion d'algèbre, qui devient, pour une part, l'étude abstraite des lois de composition régissant les éléments d'un système de nombres. La notion générale de loi de composition s'était dégagée de travaux de Gauss sur certaines formes quadratiques, de la théorie des groupes de substitution, et des travaux d'algèbre abstraite et de logique symbolique de l'école anglaise. (Nous laissons de côté la contribution de la logique au développement des mathématiques au XIXe siècle; pour l'histoire de la logique, voir le chapitre qui lui est consacré en particulier. Nous profitons de cette occasion pour dire que nous laisserons également de côté l'histoire du calcul des probabilités et des statistiques qui est évoquée dans le chapitre *Mathématiques et société*). La découverte du calcul matriciel, du calcul géométrique de Grassmann, celle des notions de corps de nombres et d'idéal exigeaient une refonte des conceptions de base et l'étude des divers types d'algèbres. C'est l'Américain B. Peirce (1809-1880) qui amorce cette refonte. Les travaux de B. Peirce, *Linear Associative Algebra* (1881), furent poursuivis par Cayley, Sylvester, C. S. Peirce (1839-1914), Dedekind, Study, Cartan, etc.

Les géométries

Au début du XIXe siècle, sous l'influence de Monge, une partie de l'école mathématique française s'appliqua à l'étude des diverses branches de la géométrie. De France, ce renouveau gagna les autres pays d'Europe, prenant des formes variées et réagissant diversement au développement parallèle de l'algèbre et de l'analyse.

La géométrie synthétique

Essor rapide de la géométrie projective, introduction des transformations géométriques marquent le développement très brillant de la géométrie synthétique ou pure dans la première moitié du XIXe siècle. La *Géométrie descriptive* de Monge a un rôle initiateur dans ce développement, et les *Annales* de Gergonne constituent un intéressant organe de liaison.

Jean-Victor Poncelet (1788-1867), ancien élève de l'École polytechnique, définit, dans son *Traité des propriétés projectives des figures,* la géométrie projective comme l'étude des propriétés invariantes par projection centrale ou perspective. Ses méthodes sont l'emploi généralisé de la perspective et des sections planes, l'introduction systématique des éléments impropres ou idéaux que Lazare Carnot avait déjà utilisés dans sa *Géométrie de position* (1803) pour définir le plan projectif à partir du plan affine par adjonction des éléments idéaux (à l'infini). La transformation par polaires réciproques, ou polaire, que Poncelet introduit est généralisée sous le nom de corrélation; la symétrie entre point et droite qui y apparaît trouve une forme plus générale dans le principe de « dualité », dont Poncelet, Gergonne, Chasles, Möbius, Plücker précisent la signification, et dont l'origine se trouve dans le théorème de Desargues selon lequel les droites prolongeant les côtés de deux triangles placés en perspective dans le même plan se coupent sur la même droite (*fig.* 16).

L'emploi que fit Poncelet des transformations géométriques : projection cylindrique ou centrale, homologie, polarité, dans le but de réduire certaines propriétés à des cas plus simples, par exemple réduire les propriétés des coniques à celles du cercle, stimula l'étude des diverses sortes de transformations.

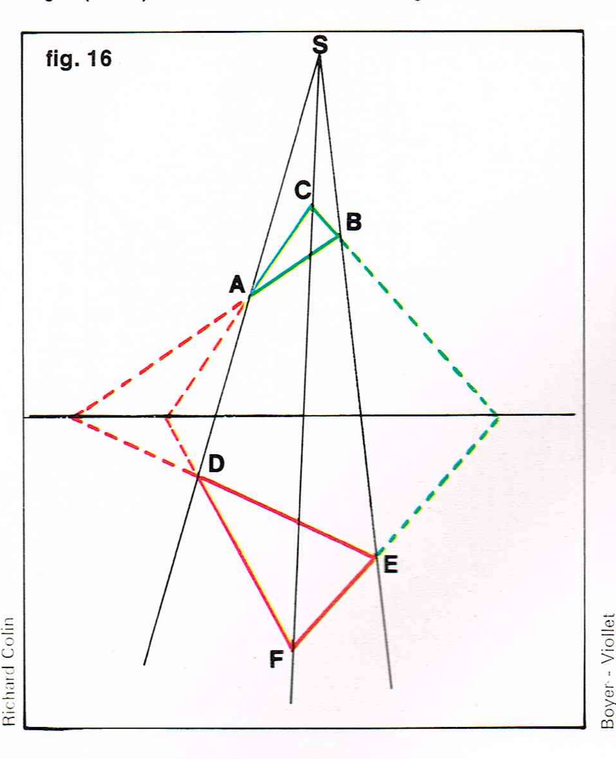
Dans sa *Systematische Entwicklung* (1832), Steiner définit dans l'espace projectif 6 formes fondamentales classées en 3 espèces et montre que l'on peut passer d'une forme à l'autre de la même espèce, pourvu qu'une condition « de projectivité », dont C. von Staudt (1847) donnera la forme générale, soit remplie. A partir de ces formes, Steiner systématise les méthodes de génération projective des figures déjà antérieurement employées dans des cas particuliers.

Chasles (1793-1880), qui devança Steiner sur certains points, contribua à la diffusion de ces méthodes. Outre de nombreux mémoires, on lui doit l'*Aperçu historique sur l'origine et le développement des méthodes géométriques* (1837), remarquable tableau de l'histoire de la géométrie, et deux traités sur les principes de la géométrie projective : *Traité de géométrie supérieure* et *Traité des sections coniques* (1852).

Staudt (1798-1867) réussit une axiomatisation de la géométrie projective, qui la débarrassa des défauts qui l'empêchaient d'être une discipline entièrement autonome : usage de notions métriques dans la définition d'éléments projectifs, justification insuffisante des éléments imaginaires, recours déguisé à la méthode des coordonnées, lourdeur de certaines démonstrations. Staudt s'efforça, en effet, de reconstituer l'ensemble de la discipline indépendamment des notions métriques d'angle et de distance, à l'aide des seuls axiomes concernant l'ordre et la position des éléments fondamentaux. Se limitant au domaine réel dans sa *Geometrie der Lage* (1847), il définit dans ses *Beiträge zur Geometrie*


Roger Viollet

▲ **Lazare Carnot** *(1753-1823); dans sa* **Géométrie de position** *(1803), il est, avec Monge, l'un des créateurs de la géométrie moderne.*



Richard Colin


Boyer - Viollet

▼ *A gauche, figure 16 : application du théorème de Desargues. A droite, le mathématicien et géomètre français* **Michel Chasles** *(1793-1880).*



▲ ***Le géomètre italien* Luigi Cremona *(1830-1903).***

der Lage (1856-1860) les éléments imaginaires comme éléments doubles d'involutions elliptiques et montre qu'ils satisfont aux axiomes fondamentaux.

Vers 1870, Klein montre la nécessité d'adjoindre aux axiomes celui de continuité; il prouve l'indépendance de la géométrie projective par rapport à l'axiome des parallèles, constate l'indémontrabilité du théorème des triangles homologiques de Desargues et du théorème de l'hexagramme de Pascal, étend la géométrie projective aux espaces à n dimensions.

La fin du siècle apporte le déclin de la géométrie synthétique au profit de sa rivale, la géométrie analytique, qui avait tiré parti des progrès de l'algèbre linéaire. De nombreux géomètres abandonnèrent le point de vue exclusif d'une géométrie pure, totalement autonome, sans craindre de profiter des ressources de l'algèbre et de l'analyse.

Les géométries non euclidiennes - Les fondements de la géométrie

Les efforts de Saccheri et de Lambert, au XVIIIe siècle, pour approfondir la signification du fameux postulat des parallèles avaient montré trois voies possibles : la voie classique fondée sur l'acceptation du postulat ou d'un équivalent; les deux autres fondées, de manières opposées, sur son rejet. Saccheri avait finalement conclu à l'absurdité de ces dernières hypothèses, tandis que Lambert, plus prudent, avait montré qu'elles se vérifient pour les figures placées sur une sphère. Dès 1792, Gauss réfléchit au problème posé par le 5e postulat, et continue de s'y intéresser toute sa vie sans rien publier. En 1799, il affirme posséder les principes d'une géométrie nouvelle, fondée sur l'existence d'une infinité de parallèles pouvant être menées par un point extérieur à une droite; il apprécie favorablement les idées de F. K. Schweikart (1819) et de Taurinus (1825-1826) qui vont dans le sens d'une géométrie non euclidienne.

Nicolas I. Lobatchevski (1792-1856), de l'université de Kazan, et János Bolyai (1802-1860), de Vienne, découvrent également, et indépendamment l'un de l'autre, la *géométrie hyperbolique* dont Gauss avait eu l'idée. Le 5e postulat est remplacé par la possibilité de mener, par un point extérieur à une droite, plusieurs parallèles à cette droite, si bien que la somme des angles d'un triangle rectiligne est inférieure à deux droits. L'originalité de Lobatchevski ne fut pas reconnue de son vivant; Gauss, qui pouvait apprécier les exposés que l'auteur fit de ses travaux, dans le *Journal de Crelle* d'abord *(Géométrie imaginaire)*, puis dans une publication berlinoise (*Geometrische Untersuchungen zur Theorie der Parallellinien* - 1840), n'exprima pas publiquement son approbation. De même, à Bolyai qui lui avait fait adresser son travail par l'intermédiaire de son père, disciple du grand mathématicien, Gauss marqua son intérêt en signalant qu'il avait lui-même depuis longtemps rencontré les mêmes idées. Bolyai, découragé, garda désormais le silence. Ce qui explique que cette géométrie resta inconnue dans ce premier temps.

Mais, en 1868, Bernhard Riemann (1826-1868) publie la thèse qu'il avait soutenue en 1854 : c'est le fameux mémoire intitulé : *Über die Hypothesen, welche der Geometrie zu Grunde liegen,* où il introduit des espaces très généraux par la donnée du carré de l'élément linéaire ds^2 et évoque le second type de géométrie non euclidienne, la géométrie « elliptique », qui nie l'infinitude de la droite et affirme que par un point extérieur à une droite il ne passe aucune parallèle à cette droite : le plan projectif ou la surface d'une sphère donnent des « modèles » de cette géométrie qui acquerra une grande importance en physique, avec la théorie de la relativité.

Les développements des géométries non euclidiennes furent ensuite fonction des travaux sur les espaces abstraits, des progrès de la théorie des groupes et de la naissance de la topologie. En 1871, Klein utilise la métrique définie par Cayley pour faire ressortir la structure d'ensemble des disciplines géométriques et interpréter la tripartition en géométrie parabolique (euclidienne), hyperbolique et elliptique. La connaissance qu'il avait de la théorie des invariants et de la théorie des groupes conduisit Klein à une belle synthèse. Dans sa célèbre dissertation de 1872, connue sous le nom de *Programme d'Erlangen,* il caractérise les différents théorèmes de géométrie et les diverses directions de recherches par les groupes de transformations qui leur correspondent. Chaque géométrie étant la théorie des invariants d'un groupe de transformations donné, les deux courants « synthétique » et « analytique » se révèlent identiques malgré la différence des langages; la géométrie euclidienne est l'étude des invariants du *groupe métrique,* la géométrie projective celle des invariants du *groupe linéaire,* la topologie celle des invariants du *groupe des transformations ponctuelles continues.* Le Programme d'Erlangen eut un grand succès et influa sur les différents domaines de la géométrie.

La diffusion des géométries non euclidiennes vint renforcer, durant la seconde moitié du XIXe siècle, l'effet produit par l'examen critique des fondements de l'analyse et les préoccupations axiomatiques en arithmétique (que nous évoquerons plus loin). Il se développa une analyse critique des principes de la géométrie classique; on explicita certains postulats implicitement admis : postulat d'Archimède mis en évidence par Stolz (et d'une certaine façon aussi par Bolzano dans des écrits qui voient le jour seulement en 1975), postulat de la continuité énoncé par Cantor et Dedekind, postulats d'ordre signalés par Gauss, Grassmann et Pasch, lequel mit l'accent sur les concepts primitifs qui permettent l'énoncé des propositions fondamentales *(postulats)* à partir desquelles se démontrent les autres propositions *(théorèmes)*.

En 1899, David Hilbert présenta une remarquable synthèse des travaux antérieurs et de ses propres recherches sur les fondements de la géométrie. Évitant le recours aux images concrètes, les *Grundlagen der Geometrie,* qui connurent rapidement plusieurs éditions (1899, 1903, 1909, etc.), introduisent trois systèmes de choses appelés *points, droites* et *plans.* Ces objets, dont la nature n'est pas précisée, entretiennent certaines relations, exprimées par 21 axiomes, classés en 5 groupes : axiomes d'appartenance, axiomes d'ordre, axiomes d'égalité ou de congruence, axiomes des parallèles et axiomes de continuité. Attentif à l'indépendance et à la non-contradiction de ces axiomes, Hilbert pense qu'ils constituent le minimum suffisant pour reconstruire tout l'édifice géométrique, à l'aide des seules règles de la logique et de l'arithmétique. Les *Grundlagen* marquent le point de départ des travaux axiomatiques dont l'importance sera un des traits essentiels de la mathématique du XXe siècle.

Le renouveau de la géométrie analytique

Il est marqué successivement par la prépondérance de l'école française de Monge, le rôle de Plücker, l'intervention de l'algèbre linéaire, l'introduction de la géométrie des droites et des espaces à n dimensions, l'essor parallèle de la géométrie algébrique et de la géométrie différentielle.

L'expression « géométrie analytique », introduite par Lacroix en 1797, fut utilisée pour la première fois dans le titre d'un ouvrage de Lefrançois en 1804. L'inscription de cette discipline dans les programmes de l'École polytechnique suscita un certain nombre de publications. Mais c'est l'école allemande qui se distingue particulièrement. Plücker dans ses *Analytisch-geometrische Entwicklungen* (2 tomes, Essen, 1828 et 1831) montre que, par l'emploi de la notation abrégée et des nouveaux systèmes de coordonnées, la géométrie analytique écarte les difficultés des calculs d'élimination et aboutit aux mêmes résultats que la géométrie pure. C'est ainsi que Plücker aboutit analytiquement au principe de dualité, précise les concepts d'équation et de coordonnées tangentielles, celui de classe d'une courbe, introduits par Möbius et Poncelet. Son *System der analytischen Geometrie* reprend l'étude et la classification des courbes algébriques, délaissées depuis le XVIIIe siècle; c'est dans cette étude qu'il introduit un principe nouveau : l'énumération des constantes, qui repose sur les célèbres formules de Plücker reliant l'ordre, la classe et les nombres des différents types de singularités (points doubles, de rebroussement, tangentes d'inflexion, tangentes stationnaires) d'une courbe de type donné.

O. Hesse (1811-1874) utilisa les déterminants et appliqua la théorie des formes algébriques et des invariants pour une présentation simple et élégante des résultats de Plücker; il établit l'équivalence entre la théorie des équations algébriques et celle des courbes

et des surfaces, donne sa forme définitive à la notation des coordonnées homogènes et introduit l'emploi du « hessien ».

En Angleterre, Cayley utilisa très largement l'algèbre linéaire dans ses études sur les transformations de coordonnées, sur les quadriques et les surfaces du 4e ordre; il généralisa les résultats de Plücker aux courbes algébriques de l'espace et aux surfaces algébriques.

L'étude des transformations projectives et des divers systèmes de coordonnées ponctuelles ou tangentielles dans l'espace à trois dimensions montra, en même temps que la symétrie des rôles joués par les points et les droites (ou les plans), la nécessité de considérer une droite tantôt comme *rayon* décrit par un point, tantôt comme *axe* autour duquel tourne un plan. Aucun des systèmes de coordonnées connus ne convenant à cette conception dualiste, Plücker conçut (en 1865) le système des coordonnées « plückériennes » (six coordonnées homogènes, l, m, n, λ, μ, ν, liées entre elles par la relation bilinéaire $l\lambda + m\mu + n\nu = 0$) pour caractériser une droite. L'originalité de cette nouvelle notation invita de nombreux mathématiciens à poursuivre l'étude de la géométrie des droites, ou géométrie réglée, et à la mettre en rapport avec celle de certaines équations aux dérivées partielles, de l'optique géométrique et de l'analyse vectorielle.

Les débuts de la géométrie algébrique

La convergence des divers travaux de géométrie renouvela, dans la seconde moitié du XIXe siècle, l'étude des courbes et des surfaces algébriques, et déboucha sur la création d'une nouvelle discipline : la géométrie algébrique, liée à la fois à la géométrie synthétique et analytique, à l'algèbre linéaire et générale et à la théorie des fonctions.

Au XVIIIe siècle, Maclaurin avait dégagé la notion de *courbe unicursale,* c'est-à-dire courbe telle que les coordonnées de son point courant sont exprimables en fonctions rationnelles d'un paramètre. Une courbe unicursale est une courbe algébrique plane qui possède le nombre maximal $N = (n-1)(n-2)/2$ de points doubles compatible avec son degré n : conique, cubique à un point double, quartique à 3 points doubles, etc.

Le théorème d'Abel (1829) sur les intégrales abéliennes éclaira cette notion en associant à chaque courbe algébrique un nombre entier $p = N-N'$, N' étant le nombre effectif de points doubles; les courbes unicursales ont donc un nombre $= 0$, ou sont de genre 0. Riemann renouvela la question en 1851 en introduisant la *surface,* dite de *Riemann,* associée à toute courbe algébrique plane. Clebsch et Poincaré complétèrent les résultats de Riemann.

Après la redécouverte de l'inversion (par Möbius, en particulier) et l'étude des transformations quadratiques (par Magnus, 1832), Jonquières donne (en 1858) le premier exemple de transformation birationnelle d'ordre quelconque, dont la théorie est édifiée par le géomètre italien Cremona (en 1863). Après quoi, on pourra concevoir les transformations birationnelles comme correspondances bijectives entre deux variétés algébriques plongées dans un espace projectif à un nombre quelconque de dimensions. La théorie des groupes permit de voir que les transformations birationnelles constituent le groupe principal de la géométrie algébrique.

L'étude des branches d'une courbe algébrique au voisinage d'un point singulier, qui constitue le centre de la géométrie algébrique, reprise par Puiseux (1850) et développée à la lumière des travaux de Riemann et de Cremona, par Noether en particulier, reçut un fondement avec le mémoire de Brill et M. Noether : *Über die algebraischen Funktionen und ihre Anwendung in der Geometrie,* publié dans les *Mathematische Annalen,* en 1874.

La géométrie infinitésimale et différentielle

La géométrie infinitésimale classique se transforme, au cours de ce siècle, en géométrie différentielle. Trois grands mathématiciens œuvrent à cette transformation : Monge, Gauss et Riemann.

En France, au début du siècle, Monge est le maître de l'école de géométrie infinitésimale, et son influence persistera, à travers ses élèves de l'École polytechnique, jusqu'à la fin du siècle, marquant encore des géomètres aussi importants que Klein, Lie ou Darboux. Le principal disciple de Monge, en cette matière, est Charles Dupin (1784-1873) qui publie *Développements de géométrie* (1813), puis *Applications de géométrie et de mécanique* (1822). Dupin introduit notamment l'indicatrice qui permet de représenter simplement la variation des rayons de courbure des sections normales en un point à une surface.

Préoccupé par divers problèmes théoriques d'astronomie, de géodésie et de cartographie (par exemple, celui de la représentation conforme d'une surface sur une autre), Gauss publie, en 1827, ses *Disquisitiones circa generales superficies curvas,* qui marquent un pas en avant par rapport à l'œuvre de Monge. Gauss introduit les coordonnées curvilignes u et v sur une surface S et donne une expression différentielle au carré de l'élément linéaire : $ds^2 = E\,du^2 + 2F\,du\,dv + G\,dv^2$ où E, F et G sont des fonctions de u et de v. Il montre que les propriétés locales de S ne dépendent pas du fait qu'elle est placée dans l'espace euclidien, mais seulement de l'élément linéaire ds. En particulier, la courbure totale en un point dépend uniquement de E, F et G, et de leurs dérivées, et demeure invariante dans les déformations des surfaces flexibles et inextensibles : c'est le « *theorema egregium* ».

Sous la triple influence de Monge, Gauss et Jacobi, d'importants travaux voient le jour, à partir de 1840, dans le *Journal de mathématiques pures et appliquées* de Liouville. C'est alors que Riemann renouvelle considérablement les principes de la discipline. Dans sa fameuse dissertation de 1854, il aborde l'étude des variétés topologiques à un nombre quelconque de dimensions, et définit par une forme quadratique positive la distance de deux points infiniment voisins : $\sum_{i,k} a_{ik}\,dx_i\,dx_k$.

▼ **Charles Dupin** ***(1784-1873), économiste et mathématicien français, fut le principal disciple de Monge (reproduction d'un cliché de Nadar).***


Archives photographiques - Paris

Riemann s'intéresse spécialement aux surfaces à courbure constante et montre comment elles permettent d'interpréter la géométrie non euclidienne plane. L'étude des géométries riemanniennes exigeait une théorie des formes différentielles quadratiques, dont Riemann lui-même amorça l'étude dans un mémoire posthume.

Entre 1864 et 1868, le géomètre italien Beltrami montra, grâce à la théorie des invariants différentiels, les liens qui unissaient les conceptions de Gauss et de Lamé et celle de Riemann. Les recherches dans cette voie aboutirent à l'analyse vectorielle et au calcul différentiel absolu.

Après 1870, les progrès de la géométrie différentielle sont marqués, d'une part, par le rôle des premières considérations topologiques, comme dans les mémoires de H. Poincaré *Sur les courbes définies par une équation différentielle* (1881-1886), qui étudient, sans intégration préalable, les propriétés des courbes intégrales d'équations différentielles, et, d'autre part, par la théorie des groupes, comme dans les œuvres de Sophus Lie : *Theorie der Transformationsgruppen* (1888-1893), *Differentialgleichungen* (1891), *Kontinuerliche Gruppen* (1893), *Berührungstransformationen* (1896).

Darboux, dans les 4 volumes de ses *Leçons sur la théorie générale des surfaces* (1887-1896), fait une remarquable synthèse de tout l'apport du siècle.

L'apparition de la topologie

Préfigurée par l'*Analysis situs* de Leibniz, cette discipline s'annonça par des problèmes célèbres, celui « des ponts de Saint-Pétersbourg » (Euler), celui des nœuds (Gauss, Listing), celui du coloriage (Möbius, de Morgan, Cayley, Tait), celui de la relation de Descartes-Euler entre les nombres de faces, d'arêtes et de sommets d'un polyèdre, avant de se préciser à travers les œuvres de Cayley, de Listing, de Möbius auquel on doit le premier exemple de surface unilatère (la bande de Möbius, 1858). C'est Riemann qui fonda cette science comme étude des propriétés invariantes par transformations bijectives continues. Les « surfaces de Riemann » font intervenir la topologie dans la théorie des fonctions de variables complexes et dans toute l'analyse (1857). Influencés par la création de la théorie des ensembles, la théorie des nombres réels et la théorie des fonctions de variables réelles, les travaux suivants portent sur les ensembles de points, sur les concepts de courbe et de domaine (Cantor, Jordan) et sur les ensembles de courbes et de fonctions.

L'analyse classique

Le développement de la physique mathématique

Il est marqué par le travail déterminant de Joseph Fourier (1768-1830). Dans ses études sur la propagation de la chaleur, commencées en 1807, présentées à l'Académie des sciences en 1811 et publiées en 1822 dans la *Théorie analytique de la chaleur,* il exprime la loi de propagation par l'équation aux dérivées partielles :

$$\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} + \frac{\partial^2 V}{\partial z^2} = a^2 \frac{\partial V}{\partial t}$$

C'est en cherchant à intégrer cette équation que Fourier représente une fonction arbitraire par la série trigonométrique qui porte, depuis, son nom :

$$f(x) = a_0 + \sum_{m=1}^{\infty} (a_m \cos m x + b_m \sin m x)$$

Fourier donne les formules qui déterminent les coefficients a_m et b_m : $a_0 = \frac{1}{2\pi}\int_0^{2\pi} f(\alpha)\, d\alpha$

$$a_m = \frac{1}{\pi}\int_0^{2\pi} f(\alpha) \cos m\alpha\, d\alpha,$$

$$b_m = \frac{1}{\pi}\int_0^{2\pi} f(\alpha) \sin m\alpha\, d\alpha,$$

posant ainsi les bases de la théorie que développera Dirichlet, et jouant un rôle décisif dans l'approfondissement des concepts fondamentaux de l'analyse.

Le renouveau de l'analyse

Bernard Bolzano, philosophe et mathématicien (1781-1848), eut une claire conscience de la nécessité de repenser avec rigueur les concepts et les définitions de base et de reconsidérer soigneusement les démonstrations pour asseoir l'analyse sur des principes irréprochables. Dès 1817, dans un mémoire consacré à la démonstration « purement analytique » d'un théorème important d'analyse (le titre complet est le suivant : *Démonstration purement analytique du théorème : entre deux valeurs quelconques qui donnent deux résultats de signes opposés se trouve au moins une racine réelle de l'équation ;* ce mémoire, assez connu en Allemagne, fut traduit en français en 1964), il prend le soin de donner une définition correcte de la notion de continuité (identique, à peu de chose près, à celle que nous avons héritée de Weierstrass, et plus rigoureuse, dans l'expression, que celle de Cauchy), en regrettant que tant de mathématiciens, « même réputés », n'aient encore à ce sujet que des concepts fort « indistincts ». Bolzano eut nombre d'idées très originales, comme celle d'avoir construit une fonction continue en tout point de son intervalle de définition, et cependant nulle part dérivable (cet exemple se trouve dans la *Funksionenlehre* éditée en 1930), mais le caractère purement logique et abstrait de son style de travail annonce plutôt les considérations qui prévaudront à la fin du siècle (avec Frege et Weierstrass notamment), et son isolement qui l'écarte des grands courants de recherches de son époque le prive du rayonnement qui lui eût permis de faire école.

Le devant de la scène est donc tout entier occupé par A. Cauchy (1789-1857), élève (d'Ampère et de Poisson en particulier) et, plus tard, professeur à l'École polytechnique ; en 1821, Cauchy publie le *Cours d'analyse* qu'il a fait à l'École, inaugurant une tradition à laquelle nous devons, ensuite, les traités de Camille Jordan (1882-1887), d'Émile Picard (1891-1896), d'Édouard Goursat (1902-1905). Les travaux de Cauchy portent sur l'analyse réelle et complexe ; à côté de découvertes originales, notamment sur l'intégrale définie, ils contiennent beaucoup de résultats connus que Cauchy reformule plus rigoureusement, dans le cadre d'un exposé systématique. C'est par leur caractère de synthèse d'ensemble que les œuvres de Cauchy connurent une si large diffusion et infléchirent le cours de l'histoire.

Dès ses premiers travaux sur les intégrales définies multiples, Cauchy remarque que l'ordre d'intégration n'est pas indifférent lorsque la fonction à intégrer devient infinie en des points intérieurs au domaine d'intégration et s'oriente vers un retour à l'ancienne définition de l'intégrale comme somme d'infiniment petits, définition dominante au XVII[e] siècle et antérieure à la contribution leibnizienne qui mit en avant, au XVIII[e] siècle, l'*intégrale indéfinie*. Dans le cours de 1821, il commence par élargir le concept de fonction numérique, en n'exigeant plus que la fonction soit définie par une expression algébrique *unique,* sans en donner toutefois un énoncé purement analytique. Mais la définition de la continuité d'une fonction contenue dans ce cours reste fortement imprégnée de géométrie.

C'est dans le résumé des *Leçons sur le calcul infinitésimal* (1823) que Cauchy donne sa nouvelle définition de l'intégrale $\int_a^b f(x)\, dx$ comme limite de la somme $\sum_{i=0}^{n-1} (x_{i+1} - x_i) f(x_i)$, avec $x_0 = a$, $x_n = b$ lorsque les intervalles $(x_{i+1} - x_i)$ tendent vers zéro. Cette définition, étendue et généralisée par Riemann dans son mémoire de 1854 : *Über die Darstellbarkeit einer Funktion durch eine trigonometrische Reihe,* est identifiée aujourd'hui sous le nom d' « intégrale de Riemann », bien qu'elle n'acquière son aspect définitif qu'avec l'exposé de Darboux (1875). Stieltjes (1894) et Henri Lebesgue (1902) donnèrent à cette notion d'intégrale définie une nouvelle extension.

Cauchy est resté également célèbre pour avoir donné plus de précision et de rigueur au *concept de convergence* d'une série ; comme il s'en explique dans la *Préface du Cours de* 1821 en des termes exemplaires, il détermina avec soin, « avant d'effectuer la sommation d'aucune série », les conditions générales de convergence des

séries, en établissant des critères dont le plus important est le fameux « critère de Cauchy » (en réalité également énoncé par Bolzano en 1817). A la suite de Cauchy, Abel et Raabe (1832), Duhamel (1839), Morgan et J. Bertrand (1842), O. Bonnet (1843), Kummer (1835), Dini (1867), enfin P. du Bois-Reymond (1873) définissent des critères de convergence de plus en plus fins. Pour les séries à termes de signes quelconques, Cauchy montre que, si une série est absolument convergente, alors elle est convergente et, complète Dirichlet en 1837, sa somme est indépendante de l'ordre de ses termes; Riemann, Abel, Dedekind, Kronecker, Weierstrass étudient également les séries non absolument convergentes. Il faut noter cependant que Cauchy ne fait pas encore la distinction entre « convergence simple » et « convergence uniforme », dont Stockes et Dirichlet auront une idée vers 1840.

Cauchy est, enfin et surtout, l'un des créateurs de la *théorie des fonctions de variable complexe*. Tandis qu'en 1821, le nombre complexe n'est pour lui qu'une « expression symbolique », il adopte, en 1825, dans son mémoire *Sur les intégrales définies prises entre des limites imaginaires,* la représentation de Gauss qui fait correspondre un point du plan à tout nombre complexe, puis, en 1849, il se rallie au point de vue de Wessel, d'Argand, de Warren, et légitime les opérations sur les complexes en les ramenant à des transformations du plan : déplacements et similitudes.

Malgré une bonne connaissance des fonctions logarithmique et exponentielle, le XVIIIe siècle n'avait pas fait d'étude systématique des fonctions de variable complexe. En 1825, Cauchy découvre que, pour une fonction définie et continue, l'intégrale $\int_a^b f(z)\, dz$ ne dépend pas du chemin le long duquel se fait l'intégration, et que, le long d'une courbe fermée sans point singulier, l'intégrale est nulle; s'il y a à l'intérieur du contour un point singulier, elle est égale à $2 \pi i R$, R étant le résidu en ce point.

Parallèlement se perfectionne la théorie des fonctions elliptiques, principalement avec Le Gendre (*Traité des fonctions elliptiques et des intégrales eulériennes,* 3 volumes de 1825 à 1832), Gauss qui ne publie rien, Abel et Jacobi qui s'aperçoivent de la nécessité de travailler dans tout le domaine de la variable complexe et d'inverser le problème en s'intéressant, non à l'intégrale elle-même, mais à la fonction inverse, de même qu'il est plus commode d'étudier la fonction *tg x* au lieu de l'intégrale $\int_o^x \frac{dx}{1+x^2}$. Abel découvre par ailleurs (en 1828) la propriété fondamentale des intégrales dites « abéliennes ». Jacobi rassemble ses résultats dans les *Fundamenta nova theoriae functionum ellipticarum* (1829) qui imposent un vocabulaire repris par Hermite et Poincaré.

Le XVIIIe siècle avait inauguré l'étude des équations différentielles et aux dérivées partielles. Cauchy établit le premier théorème d'existence d'une solution pour une équation différentielle donnée, par une méthode encore en usage aujourd'hui : elle consiste à remplacer l'équation différentielle par une équation aux différences finies, dont on fait tendre ensuite les « pas » vers zéro. En 1868, Lipschitz retrouve cette méthode en en précisant les conditions d'application, dites conditions « lipschitziennes ». Une seconde méthode, également due à Cauchy, et retrouvée dans toute sa généralité par E. Picard en 1890, est celle des « fonctions majorantes », qui repose sur le développement en série des fonctions analytiques; Cauchy a démontré la convergence des séries entières exprimant les solutions d'un système d'équations différentielles, ce que Darboux et Sophia Kowalevskaïa retrouvent plus tard par une voie plus élégante. Quant aux méthodes d'intégration de ces équations, c'est le géomètre Sophus Lie qui met de l'ordre dans la profusion des résultats obtenus par Cauchy, Jacobi, Clebsch, Liouville, grâce à sa théorie des groupes continus de transformations qui permet et d'unifier sous un seul principe les méthodes classiques, et de déduire de la nature du groupe supposé connu des résultats précis sur la nature des systèmes auxiliaires intervenant dans l'intégration.

Les travaux de Riemann et de Weierstrass

Ces deux noms dominent toute la seconde moitié du XIXe siècle et, par des méthodes différentes et complémentaires, impriment à la *théorie des fonctions analytiques* sa forme classique.

Bernhard Riemann (1826-1866), un des mathématiciens les plus profonds, fut l'élève de Gauss à Göttingen, puis de Jacobi et de Dirichlet à Berlin. En 1859, il succède à Dirichlet dans la chaire de Gauss à Göttingen, après avoir soutenu sa thèse en 1851. Celle-ci, *Grundlagen für eine allgemeine Theorie der Funktionen einer veränderlichen komplexen Grösse,* est fondamentale. Tout à fait indépendante des idées de Cauchy, elle s'inspire de la physique mathématique et de la géométrie et nous montre Riemann en disciple de Gauss, dont l'œuvre plus soignée que celle de Cauchy et cependant moins connue du grand public mathématique fut déterminante pour le progrès ultérieur de la physique et des mathématiques.

La théorie du potentiel, commencée au XVIIIe siècle, est surtout développée au XIXe siècle; l'expression même est introduite par Green qui montre, en 1843, l'invariance de l'équation de Laplace par rapport à son inversion. Riemann s'intéresse à cette théorie, guidé probablement par l'enseignement de Gauss qui avait beaucoup travaillé la question. Riemann résout le problème dit « de Dirichlet » : déterminer une intégrale par ses valeurs sur un contour fermé, en appliquant le « principe de Dirichlet », qu'avaient utilisé aussi bien Gauss que Jacobi. Après une objection de Weierstrass et les recherches de Schwarz, C. G. Neumann et Poincaré, Hilbert démontre en 1900 qu'on peut apporter la rigueur voulue à la démonstration de Riemann.

Le caractère topologique des méthodes de Riemann se marque davantage dans sa *Theorie der Abelschen Funktionen* (1857), qui fonde véritablement la théorie des fonctions algébriques, en introduisant les « surfaces de Riemann », formées de plans superposés, en nombre égal au degré d'une équation algébrique et reliés par des lignes de passage joignant les points critiques. L'influence de Riemann fut grande et immédiate, en Allemagne aussi bien qu'en France et en Italie.

Karl Weierstrass (1815-1897) fut initié aux fonctions elliptiques par Gudermann qui l'incita à se fonder sur les développements en séries entières. Après de longues années passées dans l'enseignement secondaire, on le retrouve en 1856 professeur à l'université de Berlin, où ses cours connurent beaucoup de succès. Comme il publiait peu, ses idées sont répandues par les notes de ses élèves. Ses méthodes eurent une influence si profonde et si durable que son nom évoque une coupure, le tournant décisif par lequel l'analyse classique a gagné son équilibre définitif contre les ambiguïtés de l'ancien calcul infinitésimal et que nous désignons, après F. Klein, par l'expression « arithmétisation de l'analyse ». Nul ne l'a reconnu peut-être de façon aussi complète que David Hilbert dans son très fameux article *Sur l'infini :* « En élucidant entre autres les notions de minimum, de fonction, de dérivée, il a écarté les objections que soulevait encore le calcul infinitésimal, il a nettoyé celui-ci de toutes les idées confuses sur l'infiniment grand et l'infiniment petit... Si aujourd'hui, grâce aux méthodes qui reposent sur la notion de nombre irrationnel, ou plus généralement sur celle de limite, il règne en analyse une harmonie et une certitude parfaites; et si, dans les questions les plus compliquées de la théorie des équations différentielles et intégrales, malgré les combinaisons les plus hardies et les plus diverses de toutes les formes de passage à la limite, tous les résultats se trouvent en accord, nous le devons essentiellement à l'activité scientifique de Weierstrass. »

Faisons brièvement le point des apports de Weierstrass. En analyse réelle, il a donc définitivement dissocié les concepts de continuité et de dérivabilité, en construisant un exemple fameux d'une fonction continue sur un segment et cependant dérivable en aucun des points du segment; l'exemple que Bolzano avait construit bien avant ne commence lui-même à être connu que vers cette époque précisément; il dissocie de même les concepts de convergence simple et de convergence uniforme, définit et utilise comme nous le faisons encore aujourd'hui la valeur absolue d'un nombre. Mais par-dessus tout, sa théorie purement arithmétique des



▲ *Le mathématicien* **Charles Hermite** *(1822-1901) s'illustra notamment par un mémoire célèbre,* **Sur la fonction exponentielle,** *dans lequel il démontra la transcendance du nombre* **e.**

nombres irrationnels lui permet de donner une définition rigoureuse de la continuité d'une fonction, celle que l'on enseigne encore aujourd'hui, et d'obtenir ainsi que l'arithmétique supplante définitivement la géométrie en analyse, ce que voulait précisément dire Klein en parlant d' « arithmétisation de l'analyse ».

Dans le domaine de l'analyse complexe, Weierstrass et, indépendamment de lui, Méray définissent la fonction par un développement en série entière au voisinage d'un point, et la déterminent ensuite de proche en proche, par un *prolongement analytique ;* les « fonctions transcendantes entières » correspondent aux séries convergentes dans tout le plan, tandis qu'au voisinage d'un point singulier, une fonction uniforme peut s'approcher de n'importe quelle valeur fixée à l'avance.

La théorie des fonctions elliptiques atteint, avec Weierstrass, son point culminant ; elle est aussitôt adoptée par Schwarz, successeur de Weierstrass à Berlin, Halphen dans son *Traité des fonctions elliptiques* (1886-1891), et par C. Jordan dans son *Cours d'analyse ;* la théorie des fonctions automorphes de Klein et de Poincaré et la solution du problème d'uniformisation lui doivent beaucoup.

La théorie des nombres

Au début du XIXe siècle, Adrien-Marie Le Gendre et surtout Gauss déterminent l'orientation des recherches en théorie des nombres. Le Gendre, dans les versions successives d'un *Essai sur la théorie des nombres,* s'appuie essentiellement sur la théorie des fractions continues dont Lagrange avait montré la fécondité. Plus fondamentales sont les *Disquisitiones arithmeticae* (1801) de Gauss, œuvre de jeunesse mais dont la rigueur est telle que, jusque vers le milieu du siècle, rien de comparable ne sera fait ni en théorie des fonctions ni en géométrie. Avec l'algèbre pure, cette théorie des nombres inaugure les mathématiques modernes telles que les conçoit le XXe siècle.

Gauss introduit la *notion* capitale *de congruence,* extension de la notion d'égalité et premier exemple des classes d'équivalence dont on sait le rôle si important aujourd'hui. « Si un nombre a divise la différence des nombres b et c, b et c sont dits *congrus* suivant a, sinon *incongrus.* a s'appellera le module » ; tels sont les termes propres de Gauss, qui poursuit : « Chacun des nombres b et c est dit *résidu* de l'autre dans le premier cas, et non résidu dans le second. » Gauss démontre rigoureusement la loi de réciprocité des résidus quadratiques, étudiée par Euler et Le Gendre.

Gauss complète les recherches de Lagrange et de Le Gendre sur les formes quadratiques : $(a, b, c) = ax^2 + bxy + cy^2$; il cherche à déterminer les nombres k représentables par une forme de ce type quand x et y sont premiers entre eux ; toutes les formes d'un déterminant donné $b^2 - 4ac$ se distribuent par classes. Les méthodes de réduction des formes sont simplifiées par Dirichlet, tandis qu'en 1851, Hermite développe une autre méthode, celle de « réduction continue ».

Fermat avait affirmé dans ses notes l'impossibilité de l'équation en nombres entiers $x^n + y^n = z^n$ pour $n \geqslant 3$ et donnait la démonstration pour les cas $n = 3$, 4. Le Gendre et Dirichlet parviennent à le démontrer pour $n = 5$ (1825) et Lamé pour $n = 7$ (1840). Or, Gauss, dans un travail de 1832, sur les résidus biquadratiques, montre que les lois élémentaires de l'arithmétique s'étendent aux entiers complexes de la forme $a + bi$, les « entiers de Gauss ». Kummer considère, plus généralement, des entiers : $a_0 + a_1r + a_2r^2 + ... + a_{n-1}\, r^{n-1}$, où les a_i sont des entiers relatifs et r une racine primitive de l'équation $r^n = 1$. Kummer arrive alors à démontrer le « grand théorème de Fermat » mais en ayant, à tort, généralisé la théorie classique de la décomposition en facteurs premiers à ses nouveaux nombres. Averti de son erreur par Dirichlet, il amende sa démonstration en inventant ses « nombres idéaux » qui donneront à Dedekind l'idée de sa théorie des idéaux, concept central dans l'étude des corps de nombres algébriques.

Si le grand théorème de Fermat a orienté la théorie des nombres vers l'extension de la notion de nombre entier, en la mettant en rapport avec l'algèbre, d'autres problèmes la tournent vers la théorie des fonctions analytiques ; ainsi le problème de la partition des nombres posé par Euler (qui fait appel aux séries entières), le problème de la répartition asymptotique des nombres premiers étudié par Le Gendre, le problème de Waring : déterminer le nombre de représentations d'un nombre n comme somme de puissances k-ièmes positives, qui fait appel à la théorie des formes, etc. Ces problèmes font intervenir un certain type de fonctions analytiques dont la plus célèbre est la *fonction zéta* de Riemann.

Pour achever, rappelons que c'est au XIXe siècle que se fit une dissociation précise entre nombres algébriques et nombres transcendants, c'est-à-dire qui ne peuvent être racines d'aucune équation algébrique à coefficients rationnels. En 1844, Liouville prouva de manière générale l'existence de nombres non algébriques ; Hermite démontra la transcendance du nombre e en 1872, et Lindemann celle du nombre π en 1882.

Conclusion

Ainsi, le XIXe siècle est avant tout le siècle où les mathématiques explosent en multiples branches diversifiées et spécialisées. Gauss, Poincaré ou Hilbert sont les derniers génies universels, tandis que commence à croître la race des experts et la nécessité pour chacun de travailler sur un ensemble de problèmes souvent restreint et impliquant un appareillage technique tel qu'il n'est que rarement accessible aux autres mathématiciens non spécialistes. Cependant, en 1893, F. Klein pensait que les concepts de groupe, de transformation linéaire et d'invariant permettaient de compenser l'extrême division des disciplines mathématiques par l'accès à un point de vue beaucoup plus général et abstrait d'où l'on aperçoit ce que Bourbaki nommera plus tard, dans son article bien connu, « l'architecture des mathématiques ».

Le XIXe siècle est, ensuite, le grand siècle de la *rigueur,* le XVIIIe ayant été sans doute celui de l'ingéniosité. En s'attachant à chercher des démonstrations même pour les résultats bien connus et en soumettant le concept de démonstration lui-même à l'examen, les mathématiciens du XIXe siècle retrouvent pour ainsi dire l'inspiration euclidienne.

Nous avons essayé de montrer que ce phénomène a surgi au confluent d'efforts aussi divers que ceux de Bolzano, Cauchy, ou Hamilton dont les quaternions, en ne vérifiant pas la loi de commutativité, ont remis en question le caractère « naturel » ou « évident » des lois de l'arithmétique en général, et au croisement de concepts aussi significatifs que ceux de *géométrie non* euclidienne, de géométrie à n dimensions, de corps *quelconque* de nombres (complexes, hypercomplexes...), de fonctions *pathologiques,* qui conduisent à repenser à neuf l'idée de *vérité* des mathématiques. Du point de vue technique, peut-être est-ce la théorie des fonctions de variable complexe qui fut la création la plus féconde. Mais du point de vue théorique, la découverte des géométries non euclidiennes apporta un changement capital et décisif dans la conception qu'on se faisait de la nature des mathématiques. Si l'on rappelle encore une fois que c'est au XIXe siècle que les concepts de nombre irrationnel d'abord, de continuité d'une fonction, de dérivée, d'intégrale ont reçu une définition rigoureuse, on reconnaîtra la composante principale de cette conception de la rigueur selon laquelle il fallait reconstruire toutes les mathématiques sur la base d'une construction rigoureuse des *nombres.* Ce qui a conduit certains (Frege ou Hilbert, par exemple) à réfléchir sur les *fondements logiques* des mathématiques, et nous avons là la deuxième composante qui s'épanouira dans une discipline alors naissante : la logique mathématique. Si la méthode axiomatique n'a pas encore imposé son point de vue sur la nature des concepts mathématiques, il n'est plus question pour le mathématicien, en tout cas, de « lire dans le grand livre de la Nature » : on a cessé de croire que les constructions mathématiques ont nécessairement un homologue dans la réalité. Nous n'en donnerons pour confirmation que ce texte de Gauss, le premier mathématicien du siècle, et un des plus grands, qui écrivait dès 1811, dans une lettre à Bessel (du 21 novembre), qu' « on ne devrait jamais oublier que les fonctions, comme toutes les autres constructions mathématiques, ne sont que notre propre création, et que, lorsque la définition qu'on commence par donner cesse d'avoir un sens, on devrait non pas s'interroger sur ce qu'elle est, mais chercher ce qu'il est convenable d'admettre pour qu'elle garde son sens ».